DNA mixtures interpretation - A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples

Manuscript Details

Manuscript number	FSIGEN_2018_145_R1
Title	DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples.
Article type	Research Paper

Abstract

The present study investigated the capabilities and performances of semi-continuous and fully-continuous probabilistic approaches to DNA mixtures interpretation, particularly when dealing with Low-Template DNA mixtures. Five statistical interpretation software, such as Lab Retriever and LRmix Studio – involving semi-continuous algorithms – and DNA•VIEW®, EuroForMix and STRmixTM – employing fully-continuous formulas – were employed to calculate likelihood ratio, comparing the prosecution and the defense hypotheses relative to a series of on-purpose prepared DNA mixtures that respectively contained 2 and 3 known contributors. National Institute of Standards and Technologies (NIST) certified templates were used for samples set up, which contained different DNA amounts for each contributor. 2-person mixtures have been prepared with proportions equal to 1:1, 19:1 and 1:19 in terms of DNA concentration. Conversely, three person mixtures were constituted by proportions equal to 20:9:1, 8:1:1, 6:3:1 and 1:1:1 in terms of DNA concentration. Furthermore, 8 equally-proportioned 3-person mixtures were prepared by means of scalar dilutions starting from an overall amount of 0.500 ng, then ranging up to DNA samples with concentrations equal to 0.004 ng (i.e. Low-Template DNA). DNA mixtures were set up in triplicate and amplified with 7 DNA amplification kits (i.e. GlobalFiler PCR Amplification Kit, NGM SElect PCR Amplification Kit, MiniFiler PCR Amplification Kit, Power Plex Fusion, PowerPlex 6C Matrix System, Power Plex ESI 17 Fast and Power Plex ESX 17 Fast) in order to evaluate whether the selection of a certain kit might represent a bias factor, capable of altering the whole interpretation process. Multi-software approach helped us to highlight any trend in the likelihood ratio results provided by semi- and fully-continuous software. As a matter of fact, fully-continuous computations provided different results in terms of degrees of magnitude of the likelihood ratio values with respect to the ones from the semi-continuous approach, regardless of the amplification kit that was utilized.

Keywords	DNA mixture interpretation; Low-Template DNA; semi-continuous model; fully-continuous model; likelihood ratio.
Taxonomy	Criminal Casework, DNA Polymorphism
Manuscript category	Regular Paper

Corresponding Author	Eugenio Alladio
Corresponding Author's Institution	Università degli Studi di Torino
Order of Authors	Eugenio Alladio, Monica Omedei, Selena Cisana, Giuseppina D'Amico, denise caneparo, Marco Vincenti, PAOLO GAROFANO

Suggested reviewers	Corina Benschop, Daniele Podini, Bruce Budowle

Submission Files Included in this PDF

File Name [File Type]

Cover Letter.docx [Cover Letter]

Response to reviewers.docx [Response to Reviewers (without Author Details)]

Highlights.docx [Highlights]

Title page.docx [Title Page (with Author Details)]

Manuscript with revision.docx [Manuscript (without Author Details)]

Figure 1.tif [Figure]

Figure 2.tif [Figure]

Figure 3.tif [Figure]

Figure 4.tif [Figure]

Table 1.docx [Table]

Table 2.docx [Table]

Authors Agreement.docx [Author Agreement]

Supplementary Material.docx [e-Component]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:
Data will be made available on request

Università degli Studi di Torino
Dipartimento di Chimica
Via P. Giuria, 7 10125 Torino Italy

Eugenio Alladio, PhD
phone: +39 3460171979
fax: +39 0116705249
e-mail: ealladio@unito.it

Torino, April 9th, 2018

Prof. A. Carracedo
Editor-in-Chief – **Forensic Science International: Genetics**

Dear Editor-in-Chief,
This letter accompanies submission to Forensic Science International: Genetics of a manuscript entitled: "*DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples* ".
The authors are:
Eugenio Alladio, Monica Omedei, Giuseppina D'Amico, Denise Caneparo, Marco Vincenti and Paolo Garofano.

Please address all correspondence to:
Eugenio Alladio, PhD
Università degli Studi di Torino, Dipartimento di Chimica
Via Pietro Giuria 7 - 10125 Torino, Italy
Mobile: +39 3460171979 Phone: +39 0116705255
E-mail: ealladio@unito.it

The present study investigated the capabilities and performances of semi-continuous and fully-continuous probabilistic approaches to DNA mixtures interpretation, particularly when dealing with Low-Template DNA mixtures. Five statistical interpretation software, such as Lab Retriever and LRmix Studio – involving semi-continuous algorithms – and DNA•VIEW®, EuroForMix and STRmixTM – employing fully-continuous formulas – were employed to calculate likelihood ratio, comparing the prosecution and the defense hypotheses relative to a series of *ad-hoc* prepared DNA mixtures that respectively contained 2 and 3 known contributors, in different proportions. Furthermore, eight equally-proportioned 3-person mixtures were prepared by means of scalar dilutions starting from an overall amount of 0.500 ng, then ranging up to DNA samples with concentrations equal to 0.004 ng (i.e. Low-Template DNA). All samples were performed in triplicate, then amplified by seven DNA amplification kits (i.e. GlobalFiler PCR Amplification Kit, NGM SElect PCR Amplification Kit, MiniFiler PCR Amplification Kit, Power Plex Fusion, PowerPlex 6C Matrix System, Power Plex ESI 17 Fast and Power Plex ESX 17 Fast) in order to evaluate whether the selection of a certain kit might represent a bias factor, capable of altering the whole interpretation process.

Novelty statement: This work is new and original and is not under consideration elsewhere. In comparison with the existing literature, the present study represents a proper interpretation approach that fulfil the extreme caution that is demanded in this forensic field, especially when Low-Template DNA and complex mixtures have to be interpreted. In particular, multi-software evaluations of a priori known 2-person and 3-person DNA mixtures prepared in our laboratory allowed us to compare the performance of both semi- and fully-continuous approaches. Log(LR) results provided by the tested fully-continuous software (i.e. DNA•VIEW®, EuroForMix and STRmixTM) turned always significantly higher than the ones calculated by the employed semi-continuous software (i.e. Lab Retriever and

Università degli Studi di Torino
Dipartimento di Chimica
Via P. Giuria, 7 10125 Torino Italy

chimica

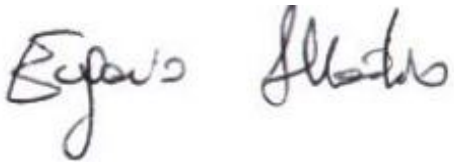
Eugenio Alladio, PhD
phone: +39 3460171979
fax: +39 0116705249
e-mail: ealladio@unito.it

LRmix Studio). Obviously, the evaluations relative to these DNA mixtures just represent a proof-of-concept that fully-continuous model seems to be the most suitable bio-statistical methodology to be performed by analysts when Low-Template DNA mixtures have to be interpreted from a probabilistic point-of-view.

Thank you for considering the paper for Forensic Science International: Genetics.

Yours faithfully,

Eugenio Alladio, PhD



Response to reviewers

Reviewer #1:

- *This paper reports the interpretation of (I think) 15 mixtures that vary in mixture ratio and template, amplified in 7 kits, and I think with two strategies for setting AT. These are interpreted in five software packages. Think this is the largest intersoftware comparison yet reported. The work concentrates on the true donor tests and does no false donor tests. This is an omission.*

  According to the comment made by the reviewer #1, the evaluation of a false donor was included within the text. Results are reported within the Supplementary Material within the Figures S6 and S8. New calculations were made by using a false donor DNA profile provided by NIST (i.e. NIST F reference samples). The results relative to such calculations have been added to the Supplementary Material. The following statement was added within the Materials and Method section "*The DNA profile from NIST F reference sample was used, conversely, as a known non contributor in order to perform false donor tests for all the tested software, too.*"
- *The conclusion seem supported EXCEPT this work really invalidates the "statistic consensus approach" which maximizes the false indication of exclusion rate. This paper gives very good evidence not to use it. The only way that I can see to support it would be to show a consequential reduction in the false inclusion rate. But this has not been investigated. It may be necessary to ask the authors to draw this conclusion.*

  In our opinion, the use of a "statistic consensus approach" seems to be useful because experts in courtrooms might show LR results from different software involving different algorithms and probabilistic approaches. Therefore, we believe that a very conservative approach might be employed in order to avoid false inclusion that could lead to unwanted legal consequences. For this reason, due to the complexity and the relevance of the task, we believe that a statistic redundancy in the calculations is preferable at the current stage. The following paper was added as reference in order to support this concept: *"Randles M., Lamb D., Odat E., Taleb-Bendiab A., Distributed redundancy and robustness in complex systems, J. Comput. Syst. Sci. 77 (2011) 293–304. doi:10.1016/j.jcss.2010.01.008"*. Moreover, non-contributor tests were also performed, as suggested by reviewer #2 too, for the three employed reference samples (i.e. NIST A, NIST B and NIST C). These results are included into the Supplementary Material. In order to reply to the correct evaluation of the reviewers, the following statements have been added within the conclusions: "Despite this fact, extreme caution has to be used when interpreting LT-DNA mixture, especially when different algorithms and probabilistic approaches are used and compared by forensic experts. Consequently, in our opinion, a conservative approach might be employed at the current stage in order to avoid false conclusions that could eventually lead to unwanted severe legal consequences. For this reason, due to the complexity and the relevance of the task, a statistic redundancy [51] in the calculations might be useful (i.e. like our adopted "statistic consensus approach").

- *Line numbers, as required in authors' instructions, would have been helpful. "Formatting requirements … Please ensure your paper has consecutive line numbering - this is an essential peer review requirement."*
  Line numbers have been added to the text, as suggested by reviewer #1.

- *drop-out rate [6,9,18–23], stutters rate. fully-continuous approaches. The binary model[1] was the first one employed by the forensic community for DNA mixture interpretation but, over the time, it was gradually replaced. turned to be not suitable since it This method does not take into account neither the instrumental stochastic effects (i.e. drop-in and drop-out –evaluated by the semi-continuous models), nor the peak heights of the detected alleles (evaluated by the fully-continuous models).*
  had to deal with bigger larger amounts of data
  traditional binary approach. For this reasons, semi-continuous [32,33] and fully continuous models [34,35] have largely replaced by far the binary one, and
  The corrections suggested by reviewer #1 have been made.

- *For this reason, in the previous years semi-continuous models (add LikeLTD [2, 3] reference to LRmix and Labretreiver please) have been largely employed by forensic analysts. since There are actually several open-source software available in literature such as, and methodologies in order to avoid misinterpretations or wrong applications of such algorithms*
  As correctly remarked by reviewer #1, LikeLTD references were added to the manuscript.

- *several mixtures were ad hoc (this is not a standard English usage.) prepared and composed by two and three known contributors mixed in different proportions. How about: mixtures of two and three known contributors mixed in different proportions were prepared.*
  As correctly remarked by reviewer #1, a correction has been made within the whole text.

- *Thanks to This experimental plan, it had been possible to even allowed the assessment of whether the performances of the examined software turned to be was influenced or not by the adopted DNA amplification kits.*
  The correction suggested by reviewer #1 has been made.

- *concentrations of DNA (i.e. up to (would this be down to?) proper LT-DNA samples). A real LT-DNA case is also discussed. Moreover, starting from such quantitative, An equally proportioned 1:1:1 mixture containing NIST samples A, B and C was serially diluted in order to obtain several further samples.*
  The corrections suggested by reviewer #1 have been made.

- *and Scientific Working Group on DNA Analysis Methods (SWGDAM) (what is the right reference for this? Maybe [4])requirements.*
  As suggested by reviewer #1 a reference (website) was added to the text.

- *concentrations Lab Retriever and LRmix Studio are open-source and free of charge software performing a semi-continuous approach to DNA mixtures.*
  Markov Chain Monte Carlo (MCMC) approach for fully-continuous DNA mixtures interpretation [29,41,45], developed by Taylor, Bright and Buckleton.
  All the described software were exploited with the aim of used to calculate LR values relative to each one of NIST samples used as known contributor and included into our prepared DNA mixtures.
  All the corrections suggested by reviewer #1 have been made.

- *A validated drop-in value equal to 0.05 (units please, maybe drop-in perlocus, if so this is VERY high).*
  A 0.05 global drop-in rate was observed and calculated during our validation studies. This values was also confirmed during our accreditation process and *strvalidator* software, too. The term *"global"* was added within the text.

- *Finally, NIST U.S. population dataset (note that this is the uncorrected database[5]) was adopted as reference database in all LR computations [46].*
  The database that was used in our study is the one that was revised in July 2017, according to https://strbase.nist.gov/NISTpop.htm. This note was added within its reference for clarification, as suggested by reviewer #1.

- *As it can be seen, LR results provided by both semi-continuous models turned very were similar or identical.*
  results provided by fully-continuous models proved similar and convergent to one another, with slightly higher within-software differences (
  Since log(LR) results turned convergent (convergent is not a standard English usage, you could define it early or maybe just use similar) among the tested approaches
  be seen, LRFC results turned were always higher than the ones provided by fullysemi-continuous modelling (LRSC) for both F6C and GF
  Similar increases in the log(LR) results from fully-continuous approach were observed, amplification kits. However, the FC results turned were always higher than the SC ones, regardless
  All the corrections suggested by reviewer #1 have been made.

- *In this case, probabilistic interpretation of 0.004 ng DNA mixture provided log(LR) values lower than zero for both the biostatistical models, May I ask for a bit more here. For example, how many unmasked alleles of the trace contributor were left above AT?*
  The number of unmasked alleles of the trace contributor left above the calculated AT for the different DNA amplification kits were around 5-20%. Different values were obtained according to the DNA amplification kit under evaluation. As a consequence, inconclusive values were "correctly" expected for these DNA mixtures. This comment has been added to the text, too.

- *with the exception of FC log(LR) result relative to NIST B known contributor equal to 1.9 Moreover, a log(LR) value of -1.47 was observed for NIST C contributor in the mixture containing 0.008 ng of DNA. Once again, these observations proved a better sensitivity of the fully-continuous models in case of LT-DNA.*
  In the present paragraph section log10(LR) results relative to a forensic real casework that was evaluated in our laboratory will be discussed. In details, a Caucasian individual was charged as an alleged suspect (POI) for a series of robberies and thefts. A cap was collected on from a crime scene by police forces during their investigating activities. Since it was supposed to belong to the alleged suspect, flocked swabs Nylon® 4N6 (purchased by COPAN ITALIA S.P.A., Brescia, Italy) were applied on the visor of the cap aiming to detect biological evidences. The genetic material was recovered on different spots on the visor, then extracted from the swabs, amplified and analyzed. Before performing our "statistic consensus approach" over acquired data, differential analytical thresholds were calculated by means of ArmedXpertTM for each dye channel, following our
  mixtures obtained were biostatistically interpreted. A summary of log(LR) results are reported in
  results with respect to the ones obtained by adopting SD3 formula. More in details, uncoherent Categorically different Results differing on which side of log(LR) = 0 were observed between semi- and fully-continuous interpretations when evaluating the biological evidence as a 2-person mixture (i.e. H(p) = POI + 1 unknown individual; H(d) = 2.
  All the corrections suggested by reviewer #1 have been made.

- *However, even though fully-continuous software showed quite high log(LR) values, semi- and fully-continuous models proved uncoherent one another gave results differing on which side of log(LR) = 0. Consequently, according to our adopted "statistic consensus approach", response expressed at the end of our interpretation process was inconclusive. This approach maximizes the false indication of exclusion rate. This paper gives very good evidence not to use it. The only way that I can see to support it would be to show a consequential reduction in the false inclusion rate. But this has not been investigated.*
  As remarked by reviewer #1 and showed by the results reported in our study, fully-continuous software provide higher absolute log(LR) values with respect to the ones provided by the semi-continuous models. However, when dealing with real caseworks, our opinion is that the forensic expert has to be very conservative, especially when interpreting LT-DNA samples. Consequently, our lab developed, validated and accredited (via ISO17025 requirements) the cited "statistic consensus approach" in order to cross-validate the results provided by the different software and provide a trustful and robust interpretation for real caseworks and, particularly, for complex LT-DNA mixtures. The authors of this manuscript are aware that this is not the best interpretation process but, according to Italian rules, this approach seemed very conservative to us (and to ACCREDIA, the Italian accreditation body, too) in order to hel the expert to provide a conclusion "beyond any reasonable doubt". Further studies are already under development by our group in order to combine the likelihood ratio results provided by different software and algorithms, using different interpretation weights, but their results are still preliminary and will not be reported in this study. A note has been added within the manuscript, as suggested by reviewer #1.

- *H(d) = 3 unknown individuals). In particular, semi-continuous approaches delivered a moderately strong support to H(p), while fully-continuous models delivered an extremely strong support such hypothesis. In the present case, response emitted the result reported at the end of our interpretation process supported the prosecution hypothesis, charging the POI as an effective contributor to the biological evidence collected on the visor of the cap that was recovered on the crime scene.*
  The correction suggested by reviewer #1 has been made.

- *Please consider excising this part: At the end of the trial, suspect was convicted as guilty since further evidences incriminated him. We should have no interest in the outcome or the "other evidence." In our opinion, results reported in Table 2 represent once again a proof to the concept that fully continuous models might be more sensitive than semi-continuous ones in case of LT-DNA mixtures interpretation.*
  In the present casework, fully-continuous log(LR) values always supported the prosecution hypothesis, which was later properly verified by the investigating authorities. Please consider excising: which was later properly verified by the investigating authorities. I cannot see how investigating authorities can verify an LR.
  The cited sentences have been removed from the text, as correctly remarked by reviewer #1.

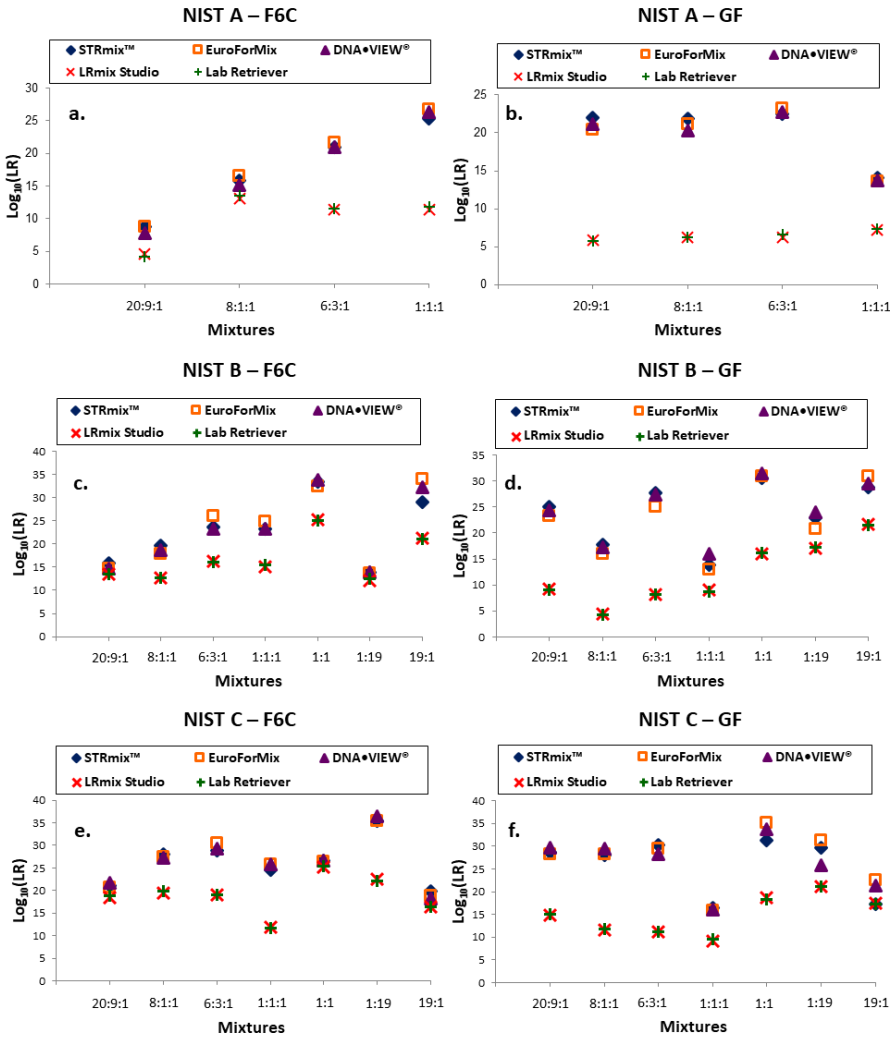
- *related to the different semi- and fully-continuous algorithms. In particular, fully-continuous software takes into account a larger amount of data and information (i.e. detected alleles plus their relative peak heights), so that higher LR values can be obtained when significant matches are observed between the investigated biological samples. Similar trends in log (LR) values were observed when several serially scalarly-diluted 3-person mixtures were investigated, too. These mixtures were ad hoc prepared in order to contain the same DNA amount for each one of the included known contributors. Nevertheless, log (LR) values provided by semi-continuous software gave false indications of exclusion turned uncoherent I cannot translate this word. In the authors mean that SC gave LRs <1 then maybe define this early as something like: false exclusionary indications to the composition of the DNA mixtures, especially in case of Low-Template DNA (i.e. mixtures showing an overall DNA concentration of 0.004 ng, 0.008 ng and 0.016 ng). Furthermore, probabilistic software behaved in a similar way, regardless of the DNA amplification kits that were employed. Even though this outcome was expected, further analyses might indicate which DNA amplification kits would turned out to be the most useful in cases of Low-Template DNA.*
  All the corrections suggested by reviewer #1 have been made.

- *Obviously, the These evaluations relative to these DNA mixtures just represent a proof-of-concept some evidence that the fully continuous model seems to be the most suitable bio-statistical methodology to be performed by analysts when Low-Template DNA mixtures have to be interpreted from a probabilistic point-of-view.*
  Further experiments (i.e. 4- and 5-person DNA mixtures[6, 7]) and their relative interpretation processes need to be performed, but in our opinion these results open the pathway towards the possibility of "weighting" LR results provided by semi- and fully-continuous models, particularly in case of LT-DNA mixtures interpretation.

All the corrections suggested by reviewer #1 have been made.

- *Table 1 and elsewhere*
  DNA Typing Kit Reference Material
  Mixtures proportion I think this is mixture ratios
  (0.500 ng)
  Table 2 Log(LR) results relative to the interpretation process performed on the DNA mixture collected on a cap recovered on a crime scene. 3SD and Min-Max represent the algorithms that were employed to evaluate the differential analytical thresholds. POI represents the suspect (i.e. the person of interest), while U stands for unknown(s) individual(s) extracted from the allele frequencies reference dataset [46].
  All the corrections suggested by reviewer #1 have been made.

- *Figure 1. Top left. I cannot see the STRmix™ results. Can we try open symbols or any other method to show them. This also applies to some of the supplementary materials.*
  All the cited figures have been modified as suggested by reviewer #1. As an example, the modified Figure 1 is reported, as follows:



- *Can I tell which component of the mixture ratio is NIST A, B or C for the three person mixtures. My guess is that it is C, B, A in that order. Please consider making the y-axis the same scale for all graphs. This also applies to some of the supplementary materials.*
  As reported in Table 1, the main contributor of the 3-person mixture is NIST A, followed by NIST B and, then, NIST C. The authors opted to leave the y-axis as it is in order not to reduce the different figures and diminish their readability. However, the y-axis can be easily modified if required by the editor, too.

- *Figure 2 Mean log(LR) values provided by semi-continuous and fully-continuous models for F6C (Figure 2a) and GF (Figure 2b) amplification kits. The dashed line represents an hypothetical situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.*
  Figure 3 Histograms displaying average log(LR) values provided by semi-continuous (SC) and fully continuous (FC) models for 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified by GF (a) and F6C (b) amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. Mean log(LR) values relative to NIST materials composing the mixtures are represented by yellow (SC) and orange (FC) histograms for NIST A contributor, light green (SC) and dark green (FC) histograms for NIST B contributor, pink (SC) and red (FC) histograms for NIST C contributor.
  All the corrections suggested by reviewer #1 have been made.

- *References:*
  All the corrections suggested by reviewer #1 have been made.

Reviewer #2:

- *The paper is interesting and it is useful to carry out comparative studies like this. However, the style of the paper and grammar are poor and needs improvement. It is quite difficult to follow. There are no line numbers or page numbers which makes referencing difficult for reviewers.*
  According to the comment of reviewer #2 and the suggestions of reviewer #1, the grammar has been riedited and, hopefully, improved. Line and page numbers have been added, too.

- *No model is "fully continuous" – because not all features are taken account of by any model. Alternative terms "qualitative" and "quantitative" models are adopted by ENFSI BPM, but terms like discrete are also acceptable. http://enfsi.eu/wp-content/uploads/2017/09/Best-Practice-Manual-for-the-internal-validation-of-probabilistic-software-to-undertake-DNA-mixture-interpretation-v1.docx.pdf.*
  Due to the fact that the terms "fully continuous" are largely reported in literature (e.g. from the paper "H. Kelly, J. Bright, J. Buckleton, J. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, Sci. Justice. 54 (2014) 66–70.

doi:10.1016/j.scijus.2013.07.003."), the authors preferred to use this way of describing the difference between the employed probabilistic approaches for DNA mixture analysis. In case also the editor would suggest to modify it, the terms "fully continuous" will be promptly replaced by a more rigorous "quantitative" model, as correctly remarked by reviewer #2.

- *Use Hp Hd rather than H(p) H(d)*
  The correction suggested by reviewer #2 has been made.

- *"several open-source software available in literature" Euroformix is open source.*
  No corrections have been made due to the fact that, in our opinion, further open source software exist like, for instance, LikeLTD and DNAmixtures. However, in case also the editor would suggest to modify this sentence, it will be promptly modified as suggested by reviewer #2.

- *"it gradually turned to be not suitable since it does not take into account neither the instrumental stochastic effects" Reference ISFG DNA commission docs.*
  No further references have been added due to the fact that, in our opinion, several papers have been already cited. However, in case also the editor would suggest to modify this reference, it will be promptly added as suggested by reviewer #2.

- *"fully-continuous approaches, that properly include peak heights' values into their algorithms, are supposed to be the most powerful methods" Who supposes this – not a scientific statement??*
  In order to fulfil the correct suggestion made by reviewer #2, the previous sentence was modified, as follows: "… are supposed to be more complex methods."

- *Note 19:1 mixtures are quite extreme and minor contributors approach the limits of interpretation.*
  19:1 mixtures have been analysed to evaluate the behaviour of the different probabilistic approach when dealing with complex and unbalanced DNA mixtures, as remarked by reviewer #2, too.

- *"slightly higher within-software differences" Do you mean between quantitative continuous software? 3 orders of magnitude? Is this a range between software across all results?*
  The different fully-continuous models provided different log(LR) values when compared one another, showing average differences of 3 orders of magnitude. However, these differences were not observed for all the tested samples and calculations, as "3 orders of magnitudes" is an indicative value. In order to better explain this concept, the sentence has been modified, as follows: " (i.e. approximatively around 3-4 degrees of magnitude, on average)".

- *You expect that the differences between software will be greater in terms of absolute magnitude when LRs are very higher. I would like to see these divergences compared to the average LR across quantitative and qualitative models. Also if all software return LRs> 1bn then there is no practical impact if the software are divergent. Divergence is more important*

when the LRs are low. In fig 1 legend it isn't clear what mixture proportion the NIST A,B,C refer to. For example does the 20:9:1 mixture refer to A:B:C?

The differences between software and the average LR across quantitative and qualitative models have been already shown in the different figures, including the ones reported in the Supplementary Material. In our opinion, the similar trends have been observed when comparing the SC and the FC models, since the log(LR) values provided by the FC software were always higher than the ones from the SC models, for all the tested mixtures. Similar results were observed when dealing with the serially scalarly-diluted 1:1:1 DNA mixtures but, as it is shown in both Figure 3 and S6, the average log(LR) values from the FC models were always higher than the ones from the SC probabilistic approach. No further graphs have been added but if also the editor would suggest to prepare them, they will be promptly added as suggested by reviewer #2 within the Supplementary Material. Finally, as it is already reported in Table 1, the main contributor of the 3-person DNA mixtures is always NIST A, followed by NIST B and, then, NIST C.

- *"Similar increase in the log(LR) results from fully-continuous approach is observed, ranging from approximately 1.4 times – for F6C – up to 1.8 times – for GF – in terms of log(LR) values, with respect to the ones provided by semi-continuous calculations." I don't see how this follows from fig 2 – please make this clearer. Surely it depends on the magnitude of the LR?*

Figure 2 shows the average log(LR) values from both SC and FC models for different DNA amplification kits. The values 1.4 and 1.8 are indicative, since they represent the slopes of the regression lines obtained when comparing the average log(LR) values of the SC models with the ones from the FC calculations. The dashed line represents the situation where all LR_{SC} are equal to LR_{FC} according to the corresponding couple of hypotheses (not the real situation). Conversely, the solid line (with intercept equal to zero) represents the average difference (or trend) between the calculated average log-likelihood ratio results provided by the two models. As a consequence, the slope is only indicative but it shows that FC models provided higher values than the SC models.

- *"In practice, SC results turned equal to -3.74 compared to its corresponding FC log(LR) value of 1.98) for NIST A and -4.73 (compared to its corresponding FC log(LR) value of 1.98) for NIST C." The authors need to consider false positive rates for the models used. A log 10 LR=1.98 would usually not be sufficient to report. I don't understand "Despite SC results delivered a very strong support to H(d), the DNA mixture under examination properly included both the contributors, as identified by FC results providing a moderate strong support to H(p)." What criteria are used to define strong support etc? log10LR=1.98 is weak surely.*

The aim of this study is not to show the consistency or the prominence of the FC models on the SC ones, but the authors want to highlight the fact the discordant LR results can be obtained, especially in case of extreme LT-DNA mixtures. The authors are aware that false positive rates should be taken into account and that a log10(LR) of 1.98 is not sufficient to be reported as an inclusion, but our aim was to show such different behaviour of the tested models. Furthermore, according to reviewer #2's comment, the sentence *"providing a moderate strong support…"* was correctly modified, as follows: "*providing a moderate support…*".

- *"These evaluations suggested that fully-continuous approaches should be adopted in case of LT-DNA interpretation." No I don't see the justification for this statement. Just because you get a bigger LR does not mean to say it is correct or preferable. This is because there is not evaluation of the effect of false positive results by carrying out non-contributor analysis which is recommended by most providers of software.*
  In our opinion, the comment made by reviewer #2 is undoubtedly correct. As a consequence, non-contributor tests were performed for all the prepared DNA mixtures. Non-contributor tests were performed for NIST A, NIST B and NIST C subjects with both LRmix Studio and EuroForMix software. The results obtained by LRmix Studio are reported in terms of boxplots in the Supplementary Material with the Figures S9a, S9b and S9c for the non-contributor tests of the subjects NIST A, NIST B and NIST C, respectively. The results provided by EuroForMix are not reported since they turned totally similar to the ones provided by LRmix Studio. Furthermore, the statement remarked by the reviewer #2 was modified, too, as follows: "*These evaluations suggested that fully-continuous approaches might provide bigger LR values in case of LT-DNA interpretation*".

- *"always supported the prosecution hypothesis, which was later properly verified by the investigating authorities." With casework you can't be sure of the ground truth so you have to be very cautious with statements like this.*
  The cited sentence has been removed as suggested by both the reviewers.

- *"Although the authors seek to make a comparative study of qualitative versus quantitative models, they omit a previous study of Bleka, Øyvind, et al. Forensic Science International: Genetics 25 (2016): 85-96 which does the same thing. Here the authors conclude "However, the main benefit of EuroForMix was with the interpretation of major/minor mixtures where the minor was evidential. Here up to 11 allele dropouts for the POI in a three-person mixture could provide probative evidence, whilst LRmix may return a much lower LR or a false negative result. The two models are expected to return similar LR results when contributors have equal mixture proportions or for mixtures of higher order"*
  As it was correctly remarked by the reviewer #2, the cited reference was added within the text.

- *The study is quite limited in that it only examines Hp: S+U vs Hd:U+U. What is the effect of conditioning with propositions like Hp: S+V vs Hd: V+U?*
  In the present study, the effect of conditioning with propositions like Hp: S+V vs Hd: V+U was not evaluated. Further studies have to be performed in order to discuss such effect, too. However, the main aim of this study deals with the evaluation of the behaviour (in terms of proof-of-concept) of SC and FC models according to different 2-person and 3-person DNA mixtures.

- *STRmix uses MCMC which means that there will be variation of LR between different runs. What is this variation and how would it impact fig 1 in comparison with the exact methods used?*
  Due to the fact that STRmix exploits MCMC methodologies, slight differences were observed in terms of LR values between consecutive STRmix runs, as correctly remarked by reviewer

#2. However, due to the fact that no significant differences were observed when STRmix calculations were randomly repeated (i.e. maximum difference lower than one order of magnitude), and due to the fact that a large number of probabilistic calculations were made, a non-comprehensive reply can be expressed for this comment. Nevertheless, the scope of the study is focused on the "proof-of-concept" relative to the evaluation of the behaviour (in terms of proof-of-concept) of SC and FC models according to different 2-person and 3-person DNA mixtures. More comprehensive validation studies dealing with STRmix approach already available in literature.

- *It isnt clear which Euroformix method is used? Bayesian? MLE? Conservative approach? All will give different results.*
  The "Continuous LR" (i.e. MLE based) methodology was employed when using EuroForMix. This information has been added within the text.

- *Remember that when assessing whether a result is good or not, does not depend upon a model giving a bigger number. We are more interested in the number of times the model returns false positive and false negative results. The authors dont address this - it would be interesting to see results of non-contributor tests.*
  As it was mentioned before, non-contributor tests were performed for all the prepared mixtures.

- *The problem with casework analysis is that you dont know the ground truth for certain - therefore you dont know if the model is behaving correctly. The grammar in the paper needs improvement.*
  As it was mentioned before, the sentence dealing the final conclusion of the caseworks were removed and, hopefully, the grammar was edited.

- Semi- and fully-continuous models to DNA mixtures interpretation are investigated
- 2- and 3-person *ad hoc* DNA mixtures analysed by multiple STR amplification kits
- Lab Retriever, LRmix Studio, DNA•VIEW®, EuroForMix and STRmix™ software were used
- LR values from fully-continuous software turned to be the highest

Title: DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples.

Authors: E. Alladio[1,2*], M. Omedei[2], G. D'Amico[2], D. Caneparo[2], M. Vincenti[1,2], P. Garofano[2,3]

Affiliations:

[1] Dipartimento di Chimica, Università degli Studi di Torino, Via P. Giuria 7, 10125 Torino, Italy.

[2] Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Regione Gonzole 10/1, 10043 Orbassano, Torino, Italy.

[3] Accademia Italiana di Scienze Forensi, Viale Regina Margherita 9/D, 42124 Reggio Emilia, Italy.

***corresponding author:**

Eugenio Alladio, PhD

Dipartimento di Chimica, Università degli Studi di Torino

Via Pietro Giuria 7,

10125 Torino, Italy

Tel.: +393460171979

Centro Regionale Antidoping e di Tossicologia "A. Bertinaria"

Regione Gonzole 10/1,

10043, Orbassano, Torino, Italy

E-mail: ealladio@unito.it; eugenio.alladio@gmail.com

Title: DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples.

Abstract:

The present study investigated the capabilities and performances of semi-continuous and fully-continuous probabilistic approaches to DNA mixtures interpretation, particularly when dealing with Low-Template DNA mixtures. Five statistical interpretation software, such as Lab Retriever and LRmix Studio – involving semi-continuous algorithms – and DNA•VIEW®, EuroForMix and STRmix™ – employing fully-continuous formulas – were employed to calculate likelihood ratio, comparing the prosecution and the defense hypotheses relative to a series of on-purpose prepared DNA mixtures that respectively contained 2 and 3 known contributors. National Institute of Standards and Technologies (NIST) certified templates were used for samples set up, which contained different DNA amounts for each contributor. 2-person mixtures have been prepared with proportions equal to 1:1, 19:1 and 1:19 in terms of DNA concentration. Conversely, three person mixtures were constituted by proportions equal to 20:9:1, 8:1:1, 6:3:1 and 1:1:1 in terms of DNA concentration. Furthermore, 8 equally-proportioned 3-person mixtures were prepared by means of scalar dilutions starting from an overall amount of 0.500 ng, then ranging up to DNA samples with concentrations equal to 0.004 ng (i.e. Low-Template DNA). DNA mixtures were set up in triplicate and amplified with 7 DNA amplification kits (i.e. GlobalFiler PCR Amplification Kit, NGM SElect PCR Amplification Kit, MiniFiler PCR Amplification Kit, Power Plex Fusion, PowerPlex 6C Matrix System, Power Plex ESI 17 Fast and Power Plex ESX 17 Fast) in order to evaluate whether the selection of a certain kit might represent a bias factor, capable of altering the whole interpretation process. Multi-software approach helped us to highlight any trend in the likelihood ratio results provided by semi- and fully-continuous software. As a matter of fact, fully-continuous computations provided different (higher) results in terms of degrees of magnitude of the likelihood ratio values with respect to the ones from the semi-continuous approach, regardless of the amplification kit that was utilized.

Keywords: DNA mixture interpretation; Low-Template DNA; semi-continuous model; fully-continuous model; likelihood ratio.

32

33

1. Introduction

The probabilistic interpretation of DNA evidences recovered on crime scenes has been for many years a largely debated and investigated issue in the field of forensic biology [1–8], especially in case of Low-Template DNA (LT-DNA) samples. There are several questions and multiple choices, indeed, that the analysts have to face when bio-statistical interpretation processes on DNA mixtures take place, such as: (i) the calculation of analytical and stochastic thresholds [9–11], (ii) the evaluation of the most probable number of contributors composing the DNA samples [12–17], (iii) the establishment of parameters such as drop-in rate, drop-out rate [6,9,18–23], stutters rate [24,25] and co-ancestry coefficient (F_{st}) [26,27], (iv) the selection of the appropriate allele frequency dataset [28]. Moreover, ~~the use of the appropriate model to be adopted on the acquired data to perform bio-statistical interpretations is essential~~bio-statistical interpretation, with the most accurate model, is essential. Briefly, there are three interpretative approaches that can be used in the sphere of DNA mixtures interpretation, which differ one another in terms of complexity, according to the input and the algorithms they take into account [29]. These methodologies are widely known as (i) binary, (ii) semi-continuous and (iii) fully-continuous approaches. The ~~B~~binary model [30] was the first one employed by the forensic community for DNA mixture interpretation but, over the time, it was gradually ~~turned~~ replaced. This method ~~to be not suitable since it~~ does not take into account neither the instrumental stochastic effects (i.e. drop-in and drop-out – evaluated by the semi-continuous models), nor the peak heights of the detected alleles (evaluated by the fully-continuous models). Due to developments of more sensitive instruments and high-throughput analyses ~~[30,31]~~[31,32], capable of evaluating very low concentrations of DNA too, analysts had to deal with ~~bigger~~ larger amounts of data and parameters that could not be managed by the traditional binary approach. For this reasons, semi-continuous [~~32,33~~33-35] and fully continuous models ~~[34,35]~~[36,37] have largely replaced by far the binary one, and nowadays they represent the gold-standard methodologies to be adopted when DNA mixture interpretations occur. However, the selection of a specific interpretation approach is not trivial, but it can be related to several factors such as, for instance, the degree of expertise that the analysts are endowed with, together with the resources of the forensic laboratories themselves. In fact, several software performing fully-continuous models are neither free of charge nor open-source, and the laboratories necessarily have to buy a license in order to use them on their own data. For this reason, in the previous years semi-continuous models have been largely employed by forensic analysts.

since there are actually several open-source software available in literature such as, for instance, Lab Retriever [33][34], and LRmix Studio (previously, LRmix [36][38]), which have been used for this study, or LikeLTD [35,39]. Another reason supporting the widespread of semi-continuous models is the fact that their algorithms and computations are more straightforward than the ones performed by fully-continuous approaches. Furthermore, their workings and results may be more easily shown and discussed in courtrooms. However, semi-continuous approach does not take into account the information regarding alleles' peak heights. As a consequence, fully-continuous approaches, that properly include peak heights' values into their algorithms, are supposed to be the most powerfulmore complex methods since they exploit the whole available information included within the acquired data. Then, fully-continuous approaches might represent a desirable solution in case of complex DNA mixture involving multiple contributors and LT-DNA. However, analysts should be well-trained before employing these methodologies in order to avoid misinterpretations or wrong applications of such algorithms. Due to these problems, our laboratory adopted a *"statistic consensus approach"* [37][40], which seems to solve several issues that traditionally arise in case of LT-DNA mixture interpretation. This approach simply compares likelihood ratio (LR) results provided by different probabilistic software, reporting only the most conservative LR value (and its correlated verbal statement [38][41]) if coherence among the tested models is observed. Otherwise, inconclusive decisions is taken into account. Even though the behaviours of semi- and fully-continuous models have been already compared in other studies [29,39–41][29,42–44], nevertheless no comprehensive guidelines have been drafted yet, describing an overall-accepted approach to be employed when dealing with complex DNA mixtures, especially in case of LT-DNA. In order to evaluate the different performances and outputs of semi- and fully-continuous models, together with the need of furtherly validating our developed *"statistic consensus approach"*, several mixtures DNA mixtures of two and three known contributors mixed in different proportions were prepared,were *ad hoc* prepared and composed by two and three known contributors mixed in different proportions, then analyzsed using 7 DNA amplification kits. 2 semi-continuous (Lab Retriever, LRmix Studio) and 3 fully-continuous (DNA•VIEW® [42][45], EuroForMix [7,35][7,37], STRmix™ [24,43][24,46]) software were employed. The provided LR results were compared, with respect to the utilized DNA amplification kit, aiming to perform a wide comparison of the cited semi-continuous and the fully-continuous software. Thanks to tThis experimental plan, it had been possible to even evaluate allowed the assessment of whether whether the performances of the examined software turned to be influenced or not by the adopted DNA amplification kits.

Furthermore, several scalar dilutions in terms of DNA amount were prepared involving a known and equally-proportioned 3-person mixture. Once again, such diluted mixtures were ~~analysed~~ analyzed with several DNA amplification kits aiming to evaluate the questioned probabilistic models with respect to decreasing concentrations of DNA (i.e. ~~up~~ down to proper LT-DNA samples). An LT-DNA real casework is also discussed~~, too~~ to test our approach.

2. Materials and methods

2.1. Sample preparation and analysis

DNA samples were all set up with Standard Reference Material® 2391c, primarily intended for use in the standardization of forensic QA (Quality Assurance) and paternity test procedures for PCR-based genetic testing. ~~NIST A, NIST B and NIST C~~ PCR-based DNA profiling standard NIST® SRM® reference samples 2391c (NIST) were selected for this experiment, and used as known contributors for 2-person and 3-person mixtures, as shown in Table 1. The DNA profile from NIST F reference sample was used, conversely, as a known non contributor in order to perform false donor tests for all the tested software, too. Different DNA proportions of NIST A, NIST B and NIST C were evaluated for both the ~~*ad hoc*~~ prepared 2- and 3-person mixtures containing known NIST contributors (mixture ratios are reported in Table 1). In particular, 2-person mixtures were prepared with contributors' ~~proportion~~ ratio of 19:1, 1:1 and 1:19, while 3-person mixtures included NIST samples mixed with the ~~proportions~~ ratios 20:9:1, 8:1:1, 6:3:1, 1:1:1. All the prepared DNA mixtures had an approximate concentration of 0.500 ng. ~~Moreover, starting from such quantitative, a~~An equally-proportioned 1:1:1 mixture (0.500 ng) containing NIST samples A, B and C was scalarly diluted in order to obtain several further samples (i.e. with lower concentration levels), such as 0.250 ng, 0.125 ng, 0.063 ng, 0.031 ng, 0.016 ng, 0.008 ng and 0.004 ng. All the samples were repeatedly amplified using the following 7 DNA amplification kits: GlobalFiler™ PCR Amplification Kit (GF), AmpFlSTR® NGM SElect™ Amplification Kit (NGM), AmpFlSTR® MiniFiler™ PCR Amplification Kit (MF) - from Thermo-Scientific, (Waltham, MA, USA) - and PowerPlex® Fusion System (F), PowerPlex® Fusion 6C System (F6C), PowerPlex® ESI 17 Fast System (ESI), PowerPlex® ESX 17 Fast System (ESX) - from Promega Corporation (Madison, WI, USA). Allele detection was performed by capillary electrophoresis (CE) on Applied Biosystems® 3500 Series Genetic Analyzer with a 36 cm 3500 Genetic Analyzer Capillary Array and POP-4™ Polymer 3500 Genetic Analyzer (Thermo Fisher Scientific) together with an Injection standard protocol 1.2 kV/15 sec.

The whole analytical methodology was internally validated following UNI CEI EN ISO/IEC 17025 and Scientific Working Group on DNA Analysis Methods (SWGDAM - http://www.swgdam.org/) requirements. Parameters such as accuracy, linearity, quantification accuracy, limit of detection (analytical threshold), limit of quantitation, mixtures deconvolution, repeatability, concordance and repeatability limit, robustness, sensitivity, decision threshold, direct amplification inhibition, stochastic threshold, specificity, species specificity, uncertainty, stutters, drop-in and drop-out were validated proving satisfactory results (not reported in the study) for all the employed DNA amplification kits. Analytical methodologies, together with the mixture interpretation approach reported in ~~[37]~~[40], were verified and accredited by ACCREDIA, the Italian body appointed for the accreditation of analytical protocols and methodologies in laboratories.

2.2. Software and LR calculations

GeneMapper® ID-X v1.4 from Thermo Fisher Scientific (Waltham, MA, USA), OSIRIS v2.7 (from http://www.ncbi.nlm.nih.gov/projects/SNP/osiris/) and ArmedXpert™ v3.0.7.999 from NicheVision Forensics LLC (Akron, OH, USA) were employed to manage the acquired raw data and filter them by applying the differential analytical thresholds calculated (i.e. specific validated analytical thresholds were observed for each dye channel). Then, data were modified on MS Excel in order to obtain suitable input formats for each bio-statistical software employed in this study, which were as follows: Lab Retriever, LRmix Studio, DNA•VIEW®, EuroForMix and STRmix™.

Lab Retriever and LRmix Studio are ~~is an~~ open-source and free of charge software performing a semi-continuous approach to DNA mixtures~~, as well as LRmix Studio~~. The first (Lab Retriever, version 2.2.1) was downloaded from the Scientific Collaboration, Innovation and Education (SCIEG) website (http://scieg.org/lab_retriever.html), developed by K. Inman, K. Lohmueller and N. Rudin. The second, LRmix Studio (version 2.1.3), is available on the website http://www.lrmixstudio.org/, developed by H. Haned and P. Gill. On the other hand, DNA•VIEW® is a commercial software involving a fully-continuous algorithm that takes into account peak height (in terms of Relative Fluorescent Units, RFU) of each detected allele. This software was developed by C.H. Brenner (http://dna-view.com/) and version 37.17 was employed to perform LR calculations based mainly on stochastic variation, incorporating dropout, drop-in, stutter and allelic stacking naturally, without the use of Markov chain Monte Carlo (MCMC) methods. Moreover, EuroForMix is an open-source and free of charge software involving a fully-continuous approach for DNA mixtures. It is one of the

first fully-continuous open-source software to be available on internet (version 1.9.3 was employed), programmed by Ø. Bleka and working in R [47] ~~[44]~~ environment (package "euroformix", R version 3.4.3 was used). In particular, it involves maximization (frequentistic) or integration (Bayesian) approaches over the likelihood function of a gamma peak height model for STR/SNP DNA data (as remarked on http://www.euroformix.com/). The "Continuous LR" mode (i.e. involving a Maximum Likelihood Estimation – MLE – methodology) was used. Finally, STRmix™ (http://strmix.esr.cri.nz/) is a commercial software involving Markov Chain Monte Carlo (MCMC) approach for fully-continuous DNA mixtures interpretation ~~[29,41,45]~~[29,44,48], developed by D. Taylor, J.A. Bright and J. Buckleton.~~J. Buckleton and his research group~~.

A free-trial version (version 2.3.06) of STRmix™ was employed in this study.

All the described software were ~~exploited with the aim of calculating~~used LR values relative to each one of NIST samples used as known contributor and included into our prepared DNA mixtures.

In case of 2-person mixtures, LR calculations were performed using the following hypotheses:

- Prosecution hypothesis Hp = Subject X_i + 1 unknown individual;
- Defence hypothesis H~~(d)~~d = 2 unknown individuals.

where X_i stands for the i-th contributor (NIST A, NIST B or NIST C) included into the mixtures under exam.

Conversely, dealing with 3-person mixtures, the following hypotheses were evaluated for the LR calculation:

- Prosecution hypothesis H~~(p)~~p = Subject X_i + 2 unknown individuals;
- Defence hypothesis H~~(d)~~d = 3 unknown individuals.

A validated and accredited overall drop-in value equal to 0.05, together with a F_{st} (theta) value of 0.01, were set in all probabilistic software. Different drop-out values, according to estimation of drop-out probability range performed by LRmix Studio ~~[36]~~ [38], were set into both Lab Retriever and LRmix Studio, but only the most conservative LR value was recorded [21]. Finally, NIST U.S. population dataset was adopted as reference database in all LR computations ~~[46]~~[49].

3. Results and discussion

3.1. LR comparison of 2-person and 3-person mixtures

LR results relative to 2-person and 3-person mixtures amplified with Fusion 6C (F6C) and GlobalFiler (GF) amplification kits are described in this section. For these two amplification kits, 2-person DNA

mixtures included individuals/certified materials labelled as NIST B and NIST C, which were mixed together in different proportions (i.e. 1:1, 1:19 and 19:1). Then, NIST A, NIST B and NIST C were used to compose four different 3-person mixtures (i.e. 20:9:1, 8:1:1, 6:3:1 and 1:1:1). In this case, NIST C was used as major contributor of DNA mixtures with ~~proportions~~ ratios of 20:9:1, 8:1:1 and 6:3:1. Moreover, NIST B certified material represented the second major contributor in the DNA mixtures with the proportions 20:9:1 and 6:3:1, while NIST A material simulated the minor contributor.

LR results of the different mixtures were evaluated comparing the prosecution hypothesis H~~(p)~~p, which included the person of interest – POI, i.e. NIST A, NIST B or NIST –, versus the defence hypothesis H~~(d)~~d, that did not include the POI, but contained unknown individuals only. Results are graphically summarized in Figure 1. In particular, base-10 logarithms were applied so that log(LR) results are reported on the y axis, while codes indicating the DNA mixtures and their relative proportions are displayed on the x axis (NIST A contributor is shown in Figure 1a-b, NIST B in Figure 1c-d and NIST C in Figure 1e-f). As it can be seen, LR results provided by both semi-continuous models ~~turned~~ were very similar or identical. This is due to the fact that Lab Retriever and LRmix Studio software utilizes similar algorithms, with slightly divergent formulas ~~[40]~~[43,44]. Furthermore, log(LR) results provided by fully-continuous models proved similar and convergent to one another, with slightly higher within-software differences (i.e. ~~up~~ approximatively around ~~to~~ 3-4 degrees of magnitude, on average), if compared to semi-continuous software.

Since log(LR) results turned ~~convergent~~ similar among the tested approaches, mean log(LR) values were calculated for both semi-continuous (LR_{SC}) and fully-continuous (LR_{FC}) algorithms. LR_{SC} and LR_{FC} results are reported in Figure 2 for F6C (Figure 2a) and GF (Figure 2b) amplification kits. As it can be seen, LR_{FC} results ~~turned~~ were always higher than the ones provided by fully-continuous modelling (LR_{SC}) for both F6C and GF. The dashed line represents ~~an~~a hypothetic situation where all LR_{SC} are equal to LR_{FC} according to the corresponding couple of hypotheses. Conversely, the solid line (with intercept equal to zero) represents the average difference (or trend) between the calculated average log-likelihood ratio results provided by the two models. Similar increases in the log(LR) results from fully-continuous approach ~~is~~ were observed, ranging from approximately 1.4 times – for F6C – up to 1.8 times – for GF – in terms of log(LR) values, with respect to the ones provided by semi-continuous calculations.

As well as F6C and GF kits, similar trends of log (LR) results were observed for all the other validated DNA amplification kits (i.e. NGM, MF, F, ESI and ESX). Results relative to these kits, together with their average log(LR) values (i.e. LR_{SC} and LR_{FC}), are reported in Figures S1-S5 in the Supplementary

Material. Furthermore, false donor test were calculated for all the tested mixtures and software by involving NIST F reference profiles. The obtained results are reported in Figure S6 in the Supplementary Material. As it can be seen, the log(LR) values provided by semi-continuous software were always lower (in terms of absolute values) than the ones provided by fully-continuous approaches.

3.2. LR variation ~~with respect to total DNA concentration~~related to DNA quantity

Eight equally-proportioned 3-person mixtures were prepared starting from an overall DNA concentration of 0.500 ng. DNA samples showing a concentration of 0.500ng, 0.250 ng, 0.125 ng, 0.063 ng, 0.031 ng, 0.016 ng, 0.008 ng and 0.004 ng were prepared including NIST A, B and C certified materials as known contributors. Once again, prosecution hypotheses H~~(p)~~p including the POI were compared to defence hypotheses H~~(d)~~d involving only unknown individuals. Base-10 logarithms were evaluated for all contributors by means of the same panel of software. Then, average log(LR) values were calculated aiming to compare the results provided by semi-continuous software (SC) with the ones provided by fully-continuous approaches (FC). As it can be seen in Figure 3a-b for GF and F6C amplification kits, respectively, log(LR) data are reported on y axis, according to employed algorithm (SC or FC), while the codes indicating the concentrations of the 1:1:1 DNA mixtures are indicated on the x axis. In particular, log(LR) results seemed to increase in correlation with the DNA concentrations of the scalarly-diluted mixtures under examination. Similar behaviours were observed for all known NIST contributors and both GF and F6C amplification kits. However, the FC results turned always higher than the SC ones, regardless of the DNA amplification kit that was adopted. This observation proved undoubtedly remarkable in case of the DNA mixtures that present very low DNA concentrations (i.e. 0.004 ng, 0.008 ng and 0.016 ng – LT-DNA). As it can be seen in Figure 3a, 0.004 ng mixture amplified using GF kit showed log(LR) values lower than zero – thus supporting the exclusionary hypothesis – when SC approach was performed involving NIST A or NIST C as POI. Otherwise, FC models provided log (LR) results higher than zero – thus supporting the inclusionary hypothesis –. In practice, SC results turned equal to -3.74 (compared to its corresponding FC log(LR) value of 1.98) for NIST A and -4.73 (compared to its corresponding FC log(LR) value of 1.98) for NIST C. Despite SC results delivered a very strong support to H~~(d)~~d, the DNA mixture under examination properly included both the contributors, as identified by FC results

providing a moderate ~~strong~~ support to H~~(p)~~p. These evaluations suggested that fully-continuous approaches ~~should be adopted~~might provide bigger log(LR) values in case of LT-DNA interpretation. This statement was corroborated by further results, such as SC log(LR) value for NIST B contributor which turned lower (i.e. 0.84, scarcely supporting H~~(p)~~p) than the one provided by FC approach (i.e. 1.78, moderately supporting H~~(d)~~d). Similar performances were observed in the cases of 0.008 ng and 0.016 ng DNA mixtures, where SC log(LR) values relative to NIST C contributor turned lower than zero, and SC log(LR) results relative to NIST A contributor turned lower (0.008 ng) or very close (0.016 ng) to zero. Moreover, log(LR) results for both SC and FC models proved always higher than zero starting from a DNA concentration of 0.031 ng, but FC values always showed higher results than those from SC approaches. Similar behaviours were observed evaluating the DNA mixtures amplified by F6C kit (Figure 3b), too. In this case, probabilistic interpretation of 0.004 ng DNA mixture provided log(LR) values lower than zero for both the biostatistical models, with the exception of FC log(LR) result relative to NIST B known contributor equal to 1.9. Moreover, a log(LR) value of -1.47 was observed for NIST C contributor in the mixture containing 0.008 ng of DNA. Once again, these observations proved a better sensitivity of the fully-continuous models in case of LT-DNA. However, the number of unmasked alleles of the trace contributors left above the calculated AT for the different DNA amplification kits were around 5-20%. Different percentage values were obtained according to the DNA amplification kit under evaluation and, therefore, inconclusive conclusion or inappropriate LR values were expected for these LT-DNA mixtures. As well as F6C and GF kits, similar performance were observed for all the DNA mixtures amplified by the other validated DNA amplification kits (i.e. NGM, MF, F, ESI and ESX). The results relative to these kits are depicted in Figure S~~6~~7 in the Supplementary Material. Furthermore, false donor tests were calculated for all the tested mixtures and software by involving NIST F reference profiles. The obtained results are reported in Figure S8 in the Supplementary Material. As it can be seen, the log(LR) values provided by semi-continuous software were, on average, similar or even higher (in terms of absolute values) than the ones provided by fully-continuous approaches when dealing with the mixture containing 0.004, 0.008, 0.016, 0.31 and 0.063 ng of DNA. Then, for DNA amounts equal to 0.125 ng or higher, the log(LR) values from the false donor calculations provided by semi-continuous software were lower (in terms of absolute values) than the ones provided by fully-continuous approaches. Non-contributor tests were also performed for all the prepared DNA mixtures. Non-contributor tests were performed for NIST A, NIST B and NIST C subjects with both LRmix Studio and EuroForMix software. The results obtained by LRmix Studio are reported in terms of boxplots in the

Supplementary Material with the Figures S9a, S9b and S9c for the non-contributor tests of the subjects NIST A, NIST B and NIST C, respectively. The results provided by EuroForMix are not reported since they turned totally similar to the ones provided by LRmix Studio.

3.3. *An intriguing real casework*

In the present paragraph log10(LR) results relative to a forensic real casework that was evaluated in our laboratory will be discussed. In details, a Caucasian individual was charged as an alleged suspect (POI) for a series of robberies and thefts. A cap was collected ~~on~~ from a crime scene by police forces during their investigating activities. Since it was supposed to belong to the alleged suspect, flocked swabs Nylon® 4N6 (purchased by COPAN ITALIA S.P.A., Brescia, Italy) were ~~applied~~ used to collect material on the visor of the cap aiming to detect biological evidences. The genetic material was recovered on different spots on the visor, then extracted from the swabs, amplified and analyzed. Before performing our *"statistic consensus approach"* over acquired data, differential analytical thresholds were calculated by means of ArmedXpert™ for each dye channel, following our validated analytical protocol. Due to the fact that several algorithms might be employed in order to calculate analytical thresholds [9,10], two different series of analytical thresholds to be applied on raw data were provided by ArmedXpert™ Analytical Threshold tool. This application involves different formulas that are labelled as SD3 (i.e. taking into account 3 standard deviations of the peak heights in terms of RFU; results are displayed in the upper part of Figure 4) and Min-Max (i.e. taking into account minima and maxima peak heights in terms of RFU; results are displayed in the lower part of Figure 4). As it can be seen in Figure 4, analytical thresholds involving SD3 algorithm showed higher thresholds than the ones obtained after the application of Min-Max algorithm. Both the series of differential analytical thresholds were applied on raw data and the DNA mixtures obtained were biostatistically interpreted. A s~~S~~ummary of log(LR) results are reported in Table 2. Same sets of H~~(p)~~p and (Hd) hypotheses including 2 or 3 contributors were evaluated and compared according to the different algorithms used in terms of analytical thresholds (3SD or Min-Max). Then, interpretative responses were stated depending on the degree of coherence among log(LR) results or, even better, between semi- and fully-continuous models. As it can be observed in Table 2, the adoption of lower RFU values (i.e. using Min-Max algorithm) provided higher log(LR) results with respect to the ones obtained by adopting SD3 formula. ~~More in details, uncoherent results were observed~~Categorically different results, i.e. differing on which side of log(LR) = 0, were observed

between semi- and fully-continuous interpretations when evaluating the biological evidence as a 2-person mixture (i.e. H~~(p)~~p = POI + 1 unknown individual; H~~(d)~~d = 2 unknown individuals). Regardless of the algorithm used to calculate analytical thresholds, both Lab Retriever and LRmix Studio provided log (LR) values lower than zero, even if very close to such value. Otherwise, DNA•VIEW®, EuroForMix and STRmix™ showed log(LR) results suggesting either moderately strong (in case of 3SD algorithm) or strong (in case of Min-Max algorithm) support to be delivered to the prosecution hypothesis that indicated an inclusionary contribution to the mixture by the POI. However, even though fully-continuous software showed quite high log(LR) values, semi- and fully-continuous models proved uncoherent one another. Despite fully-continuous software provided higher absolute log(LR) values with respect to the ones from the semi-continuous models, our lab adopted, validated and accredited (via UNI CEI EN ISO/IEC 17025 requirements) the cited *"statistic consensus approach"* in order to cross-validate the results provided by the different software and provide a trustful and robust interpretation for real caseworks and, particularly, for complex LT-DNA mixtures. Consequently, according to our adopted *"statistic consensus approach"*, *the* response expressed at the end of our interpretation process was inconclusive. ~~Furthermore~~On the other hand, coherent results were observed when evaluating the biological evidence as a 3-person mixture (i.e. H~~(p)~~p = POI + 2 unknown individuals; H~~(d)~~d = 3 unknown individuals). In particular, semi-continuous approaches delivered a moderately strong support to H~~(p)~~p, while fully-continuous models delivered an extremely strong support to such hypothesis. In the present case, ~~response emitted~~the result reported at the end of our interpretation process supported the prosecution hypothesis, charging the POI as an effective contributor to the biological evidence collected on the visor of the cap that was recovered on the crime scene. ~~At the end of the trial, suspect was convicted as guilty since further evidences incriminated him.~~ In our opinion, results reported in Table 2 represent once again a proof to the concept that fully-continuous models might be more sensitive than semi-continuous ones in case of LT-DNA mixtures interpretation. In the present casework, fully-continuous log(LR) values always supported the prosecution hypothesis., ~~which was later properly verified by the investigating authorities~~.

Conclusions

Thanks to the recent developments involving DNA extraction and typing, the interpretation of the biological evidence is gradually being more and more crucial in trials, strongly affecting the judgments expressed in courtrooms. As a consequence, extreme caution is demanded in this forensic field, especially when Low-Template DNA and complex mixtures have to be interpreted. Semi-continuous and fully-continuous approaches present different degrees of complexity in terms of comprehensibility, computation and interpretation of the outputs. Multi-software evaluations of *a priori* known 2-person and 3-person DNA mixtures prepared in our laboratory allowed us to compare the performance of both semi- and fully-continuous approaches. Log(LR) results provided by the tested fully-continuous software (i.e. DNA•VIEW®, EuroForMix and STRmix™) turned always significantly higher than the ones calculated by the employed semi-continuous software (i.e. Lab Retriever and LRmix Studio). A plausible explanation of this might be related to the different semi- and fully-continuous algorithms. In particular, fully-continuous software takes into account a larger amount of data and information (i.e. detected alleles plus their relative peak heights), so that higher LR values can be obtained when significant matches are observed between the investigated biological samples. Similar trends in log (LR) values were observed when several serially scalarly-diluted 3-person mixtures were investigated, too. These mixtures were ~~*ad hoc*~~ prepared in order to contain the same DNA amount for each one of the included known contributors. Nevertheless, log (LR) values provided by semi-continuous software gave false exclusionary indications to the composition of the DNA mixtures, especially in case of Low-Template DNA (i.e. mixtures showing an overall DNA concentration of 0.004 ng, 0.008 ng and 0.016 ng). ~~turned uncoherent to the composition of the DNA mixtures, especially in case of Low-Template DNA (i.e. mixtures showing an overall DNA concentration of 0.004 ng, 0.008 ng and 0.016 ng).~~ Furthermore, probabilistic software behaved in a similar way, regardless of the DNA amplification kits that were employed. Even though this outcome was expected, further analyses might indicate which DNA amplification kits would turned out to be the most useful in cases of Low-Template DNA. ~~Obviously, the~~These evaluations relative to these DNA mixtures just represent ~~a proof-of-concept~~some evidence that fully-continuous model seems to be the most suitable bio-statistical methodology to be performed by analysts when Low-Template DNA mixtures or major/minor mixtures, where the minor contributor is evidential [50], have to be interpreted from a probabilistic point-of-view. Furthermore, the fully-continuous model is more effective to determine the contributors' ratios. Despite this fact, extreme caution has to be used when interpreting LT-DNA mixture, especially when different algorithms and

probabilistic approaches are used and compared by forensic experts. Consequently, in our opinion, a conservative approach might be employed at the current stage in order to avoid false conclusions that could eventually lead to unwanted severe legal consequences. For this reason, due to the complexity and the relevance of the task, a statistic redundancy [51] in the calculations might be useful (i.e. like our adopted *"statistic consensus approach")*. Further studies will be performed in order to combine the likelihood ratio results provided by different software and algorithms, using different interpretation weights. ~~Further~~ Moreover, further experiments (i.e. 4- and 5-person DNA mixtures [52,53]) and their relative interpretation processes need to be performed, but in our opinion these results open the pathway towards the possibility of "weighting" LR results provided by semi- and fully-continuous models, particularly in case of LT-DNA mixtures interpretation.

Conflict of interest statement

The authors of this manuscript certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

[1] P. Gill, R. Sparkes, R. Pinchin, Interpreting simple STR mixtures using allele peak areas, Forensic Sci. Int. 91 (1998) 41–53. doi: 10.1016/S0379-0738(97)00174-6.

[2] J. Buckleton, C. Triggs, S. Walsh, Forensic DNA evidence interpretation, First Ed., Boca Raton, Florida, US, 2005.

[3] M. Bill, P. Gill, J. Curran, T. Clayton, PENDULUM—a guideline-based approach to the interpretation of STR mixtures, Forensic Sci. Int. Genet. 148 (2005) 181–189. doi:10.1016/j.forsciint.2004.06.037.

[4] B. Budowle, A.J. Onorato, T.F. Callaghan, A. Della Manna, A.M. Gross, R.A. Guerrieri, et al., Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed dna profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821. doi:10.1111/j.1556-4029.2009.01046.x.

[5] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice. 51 (2011) 204–208. doi:10.1016/j.scijus.2011.08.004.

[6] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (2012) 762–774. doi:10.1016/j.fsigen.2012.08.008.

[7] Ø. Bleka, P. Gill, Interpretation of a complex STR DNA profile using EuroForMix, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e405–e406. doi:10.1016/j.fsigss.2015.09.160.

[8] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches—Twenty years of research and development, Forensic Sci. Int. Genet. 18 (2015) 100–117. doi:10.1016/j.fsigen.2015.03.014.

[9] C.A. Rakay, J. Bregu, C.M. Grgicak, Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out, Forensic Sci. Int. Genet. 6 (2012) 723–728. doi:10.1016/j.fsigen.2012.06.012.

[10] J. Bregu, D. Conklin, E. Coronado, M. Terrill, R.W. Cotton, C.M. Grgicak, Analytical Thresholds and Sensitivity: Establishing RFU Thresholds for Forensic DNA Analysis, J. Forensic Sci. 58 (2013) 120–129. doi:10.1111/1556-4029.12008.

[11] P. Gill, R. Puch-Solis, J. Curran, The low-template-DNA (stochastic) threshold—Its determination relative to risk analysis for national DNA databases, Forensic Sci. Int. Genet. 3 (2009) 104–111. doi:10.1016/j.fsigen.2008.11.009.

[12] J.A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci. Int. Genet. 12 (2014) 208–214. doi:10.1016/j.fsigen.2014.06.009.

[13] S.L. Lauritzen, J. Mortera, Bounding the number of contributors to mixed DNA stains, Forensic Sci. Int. 130 (2002) 125–126. doi:10.1016/S0379-0738(02)00351-1.

[14] D. Taylor, J.-A. Bright, J. Buckleton, Interpreting forensic DNA profiling evidence without specifying the number of contributors, Forensic Sci. Int. Genet. 13 (2014) 269–280. doi:10.1016/j.fsigen.2014.08.014.

[15] H. Haned, L. Pène, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. 5 (2011) 281–284. doi:10.1016/j.fsigen.2010.04.005.

[16] J. Curran, P. Gill, M. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53. doi:10.1016/j.forsciint.2004.04.077.

[17] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (2007) 20–28. doi:10.1016/j.fsigen.2006.09.002.

[18] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688. doi:10.1016/j.fsigen.2012.06.002.

[19] H. Haned, T. Egeland, D. Pontier, Estimating drop-out probabilities in forensic DNA samples: a simulation approach to evaluate different models, Forensic Sci. Int. Genet. 5 (2011) 525–531. doi:10.1016/j.fsigen.2010.12.002.

[20] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, M. Prinz, T. Caragine, Likelihood ratio statistics for DNA mixtures allowing for drop-out and drop-in, Forensic Sci. Int. Genet. Suppl. Ser. 3 (2011) e240–e241. doi:10.1016/j.fsigss.2011.08.119.

[21] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Aleman, M. Aler, et al., Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54. doi:10.1016/j.fsigen.2013.10.011.

[22] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (2012) 749–761. doi:10.1016/j.fsigen.2012.08.007.

[23] R. Puch-Solis, A dropin peak height model, Forensic Sci. Int. Genet. 11 (2014) 80–84. doi:10.1016/j.fsigen.2014.02.005.

[24] J.A. Bright, D. Taylor, J. Curran, J. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304. doi:10.1016/j.fsigen.2012.11.013.

[25] J.A. Bright, J.M. Curran, Investigation into stutter ratio variability between different laboratories, Forensic Sci. Int. Genet. 13 (2014) 79–81. doi:10.1016/j.fsigen.2014.07.003.

[26] I.W. Evett, B. Weir, Interpreting DNA Evidence - Statistical Genetics for Forensic Scientists, First Ed., Sinaur Associates Inc., Sunderland, MA, USA, 1998.

[27] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140. doi:10.1016/0379-0738(94)90222-4.

[28] C.D. Steele, D.J. Balding, Choice of population database for forensic DNA profile analysis, Sci. Justice. 54 (2014) 487–493. doi:10.1016/j.scijus.2014.10.004.

[29] T.W. Bille, S.M. Weitz, M.D. Coble, J. Buckleton, J.A. Bright, Comparison of the performance of different models for the interpretation of low level mixed DNA profiles, Electrophor. 35 (2014) 3125–3133. doi:10.1002/elps.201400110.

[30] T. Clayton, J.P. Whitaker, R.L. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Science International 91 (1998) 55–70. doi: 10.1016/S0379-0738(97)00175-8.

[31] J.A. Bright, S. Neville, J.M. Curran, J.S. Buckleton, Variability of mixed DNA profiles separated on a 3130 and 3500 capillary electrophoresis instrument, Aust. J. Forensic Sci. 46 (2013) 304–312. doi:10.1080/00450618.2013.851279.

[32] A. Kirkham, J. Haley, Y. Haile, A. Grout, High-throughput analysis using AmpF l STR® Identifiler® with the Applied Biosystems 3500 xl Genetic Analyser, Forensic Sci. Int. Genet. 7 (2013) 92–97. doi:10.1016/j.fsigen.2012.07.003.

[33] H. Haned, Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268. doi:10.1016/j.fsigen.2010.03.017.

[34] K.E. Lohmueller, N. Rudin, Calculating the Weight of Evidence in Low-Template Forensic DNA Casework, J. Forensic Sci. 58 (2013) S243–S249. doi:10.1111/1556-4029.12017.

[35] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10. doi:10.1016/j.fsigen.2009.03.003.

[36] M.R. Wilson, In response to "Comparison of the performance of different models for the interpretation of low level mixed DNA profiles," Electrophor. 35 (2014) 489–494. doi:10.1080/13518040701205365.

[37] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016) 35–44. doi:10.1016/j.fsigen.2015.11.008.

[38] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263. doi:10.1016/j.fsigen.2012.11.002.

[39] D.J. Balding, C.D. Steele, likeLTD v6.0: an illustrative analysis, explanation of the model, results of validation tests and version history. <https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code>, 2015 (accessed 17.12.2015.).

[40] P. Garofano, D. Caneparo, G. D'Amico, M. Vincenti, E. Alladio, An alternative application of the consensus method to DNA typing interpretation for Low Template-DNA mixtures, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e422–e424. doi:10.1016/j.fsigss.2015.09.168.

[41] G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science - Evidential Value of Multivariate Physicochemical Data, First Ed., John Wiley & Sons, Ltd., Chicester, UK, 2014.

[42] H. Kelly, J. Bright, J. Buckleton, J. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, Sci. Justice. 54 (2014) 66–70. doi:10.1016/j.scijus.2013.07.003.

[43] J.A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, Forensic Sci. Int. Genet. 14 (2015) 125–131. doi:10.1016/j.fsigen.2014.09.019.

[44] J.A. Bright, K.E. Stevenson, J.M. Curran, J.S. Buckleton, The variability in likelihood ratios due to different mechanisms, Forensic Sci. Int. Genet. 14 (2015) 187–190. doi:10.1016/j.fsigen.2014.10.013.

[45] C. Brenner, http://dna-view.com/ (last accessed March 2nd, 2018).

[46] D. Taylor, J.A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528. doi:10.1016/j.fsigen.2013.05.011.

[47] R Development Core Team, R: A language and environment for statistical computing., (2008). http://www.r-project.org.

[48] J.A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (2013) 445–449. doi:10.1080/00450618.2013.772235.

[49] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Sci. Int. Genet. 7 (2013) e82–e83. doi:10.1016/j.fsigen.2012.12.004. (Revised in July 2017).

[50] Ø. Bleka, C.C.G. Benschop, G. Storvik, P. Gill, A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles, Forensic Sci. Int. Genet. 25 (2016) 85–96. doi: 10.1016/j.fsigen.2016.07.016.

[5~~10~~] M. Randles, D. Lamb, E. Odat, A. Taleb-Bendiab, Distributed redundancy and robustness in complex systems, J. Comput. Syst. Sci. 77 (2011) 293–304. doi:10.1016/j.jcss.2010.01.008.

[5~~21~~] J.A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R.H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M. Greer-Ritzheimer, V. Beamer, D.A. Taylor, J.S. Buckleton, Internal validation of STRmix™; A multi laboratory response to PCAST, Forensic Sci. Int. Genet. 34 (2018) 11–24. doi: 10.1016/j.fsigen.2018.01.003.

[5~~32~~] T.R. Moretti, R.S. Just, S.C. Kehl, L.E. Willis, J.S. Buckleton, J.-A. Bright, D.A. Taylor, A.J. Onorato, Internal validation of STRmix™; for the interpretation of single source and mixed DNA profiles, Forensic Science International: Genetics 29 (2017) 126-144. doi: 10.1016/j.fsigen.2017.04.004.

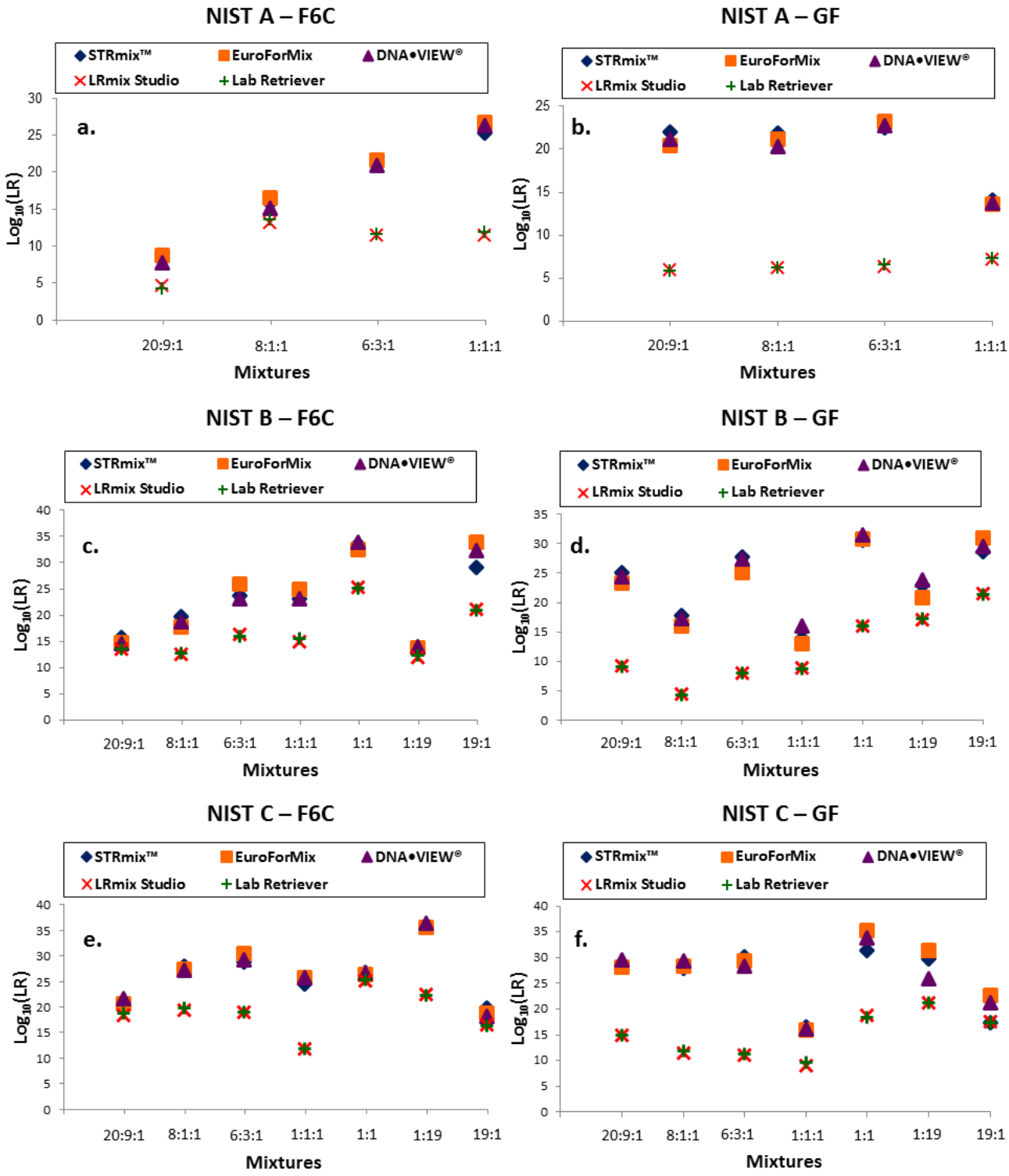
Table 1 List of NIST samples used as known contributors for DNA mixtures set up. Different proportions were evaluated both for 2- and 3-person mixtures. Referring to 2-person mixtures, different NIST reference materials were used for samples preparation according to DNA amplification kits user guide.

Known 2-person mixtures		
DNA Typing Kit	**Reference Material**	**Mixtures ~~proportion~~ratios (0.500 ng)**
ESI 17 Fast	B:C / male:male	19:1 8:1 1:1 1:19
ESX 17 Fast	A:C / female:male	
Fusion	B:C / male:male	
Fusion 6C	B:C / male:male	
GlobalFiler	B:C / male:male	
Mini Filer	A:C / female:male	
NGM SElect	A:C / female:male	

Known 3-person mixtures		
DNA Typing Kit	**Reference Material**	**Mixtures ~~proportion~~ratios (0.500 ng)**
All kits	A:B:C / Female:male:male	1:1:1 6:3:1 8:1:1 20:9:1

Table 2 Log(LR) results relative to the interpretation process performed on the DNA mixture collected on a cap recovered on a crime scene. 3SD and Min-Max represent the algorithms that were employed to evaluate the differential analytical thresholds. POI represents the suspect (i.e. the person of interest), while U stands for unknown(s) individual(s) extracted from the allele frequencies reference dataset~~Log(LR) results relative to the interpretation process performed on the DNA mixture collected on a cap recovered on a crime scene. 3SD and Min-Max represent the algorithms that were employed to evaluate the differential analytical thresholds. POI represents the suspect (i.e. the person of interest), while U stands for unknown(s) individual(s) extracted from the allele frequencies reference dataset~~ [4~~6~~7].

Analytical Threshold	**3SD**	**Min-Max**	**3SD**	**Min-Max**
Hp	Susp + 1U	Susp + 1U	Susp + 2U	Susp + 2U
Hd	2U	2U	3U	3U
Software			**Log(LR)**	
Lab Retriever	-0.36	-0.16	2.37	2.49
LRmix Studio	-0.37	-0.16	2.36	2.48
DNA●VIEW®	2.29	3.47	6.79	7.06
EuroForMix	2.33	3.60	6.25	7.12
STRmix™	2.84	3.63	7.01	7.03
Interpretative decision	**Inconclusive**	**Inconclusive**	**Support to H~~(p)~~p**	**Support to H~~(p)~~p**



571

572 **Figure 1** Log(LR) results relative to 2-person and 3-person mixtures that were amplified with Fusion 6C (F6C,
573 figures 1a, 1c, 1e) and GlobalFiler (GF, figures 1b, 1d, 1f) DNA amplification kits. Log(LR) results of the known
574 contributor labelled as NIST A (a-b), NIST B (c-d) and NIST C (e-f) are shown. In particular, log(LR) values are
575 represented by: blue diamonds for STRmix™, orange squares for EuroffForMmix, purple triangles for
576 DNA•VIEW®, red addition marks for LRmix Studio and green crosses for Lab Retriever.

577

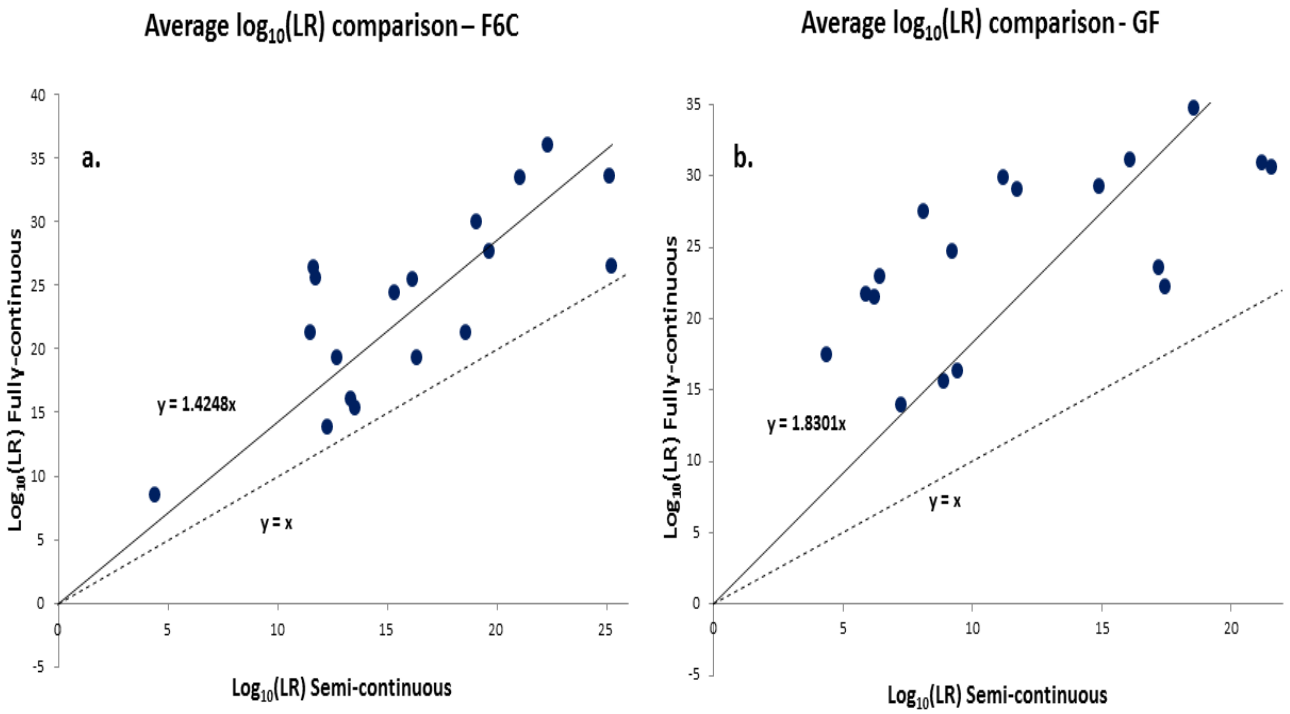
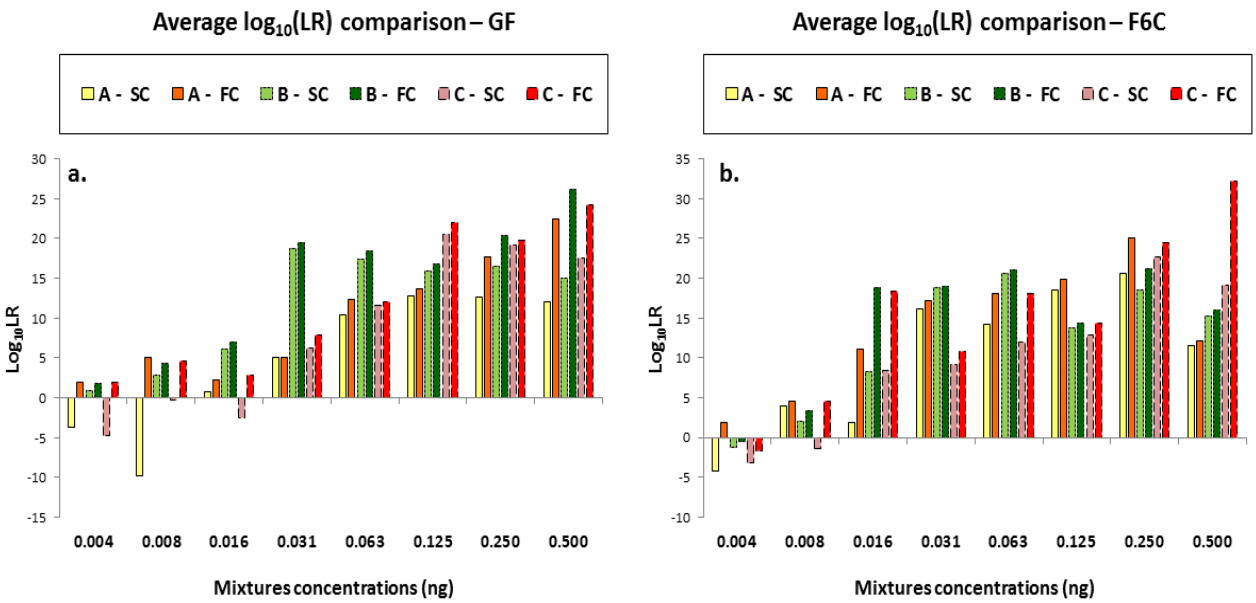


Figure 2 Mean log(LR) values provided by semi-continuous and fully-continuous models for F6C (Figure 2a) and GF (Figure 2b) amplification kits. The dashed line represents an hypothetical situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.



586

587 **Figure 3** Histograms displaying average log(LR) values provided by semi-continuous (SC) and fully-continuous
588 (FC) models for 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified by GF (a) and F6C
589 (b) amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. Mean
590 log(LR) values relative to NIST materials composing the mixtures are represented by yellow (SC) and orange
591 (FC) histograms for NIST A contributor, light green (SC) and dark green (FC) histograms for NIST B contributor,
592 pink (SC) and red (FC) histograms for NIST C contributor.

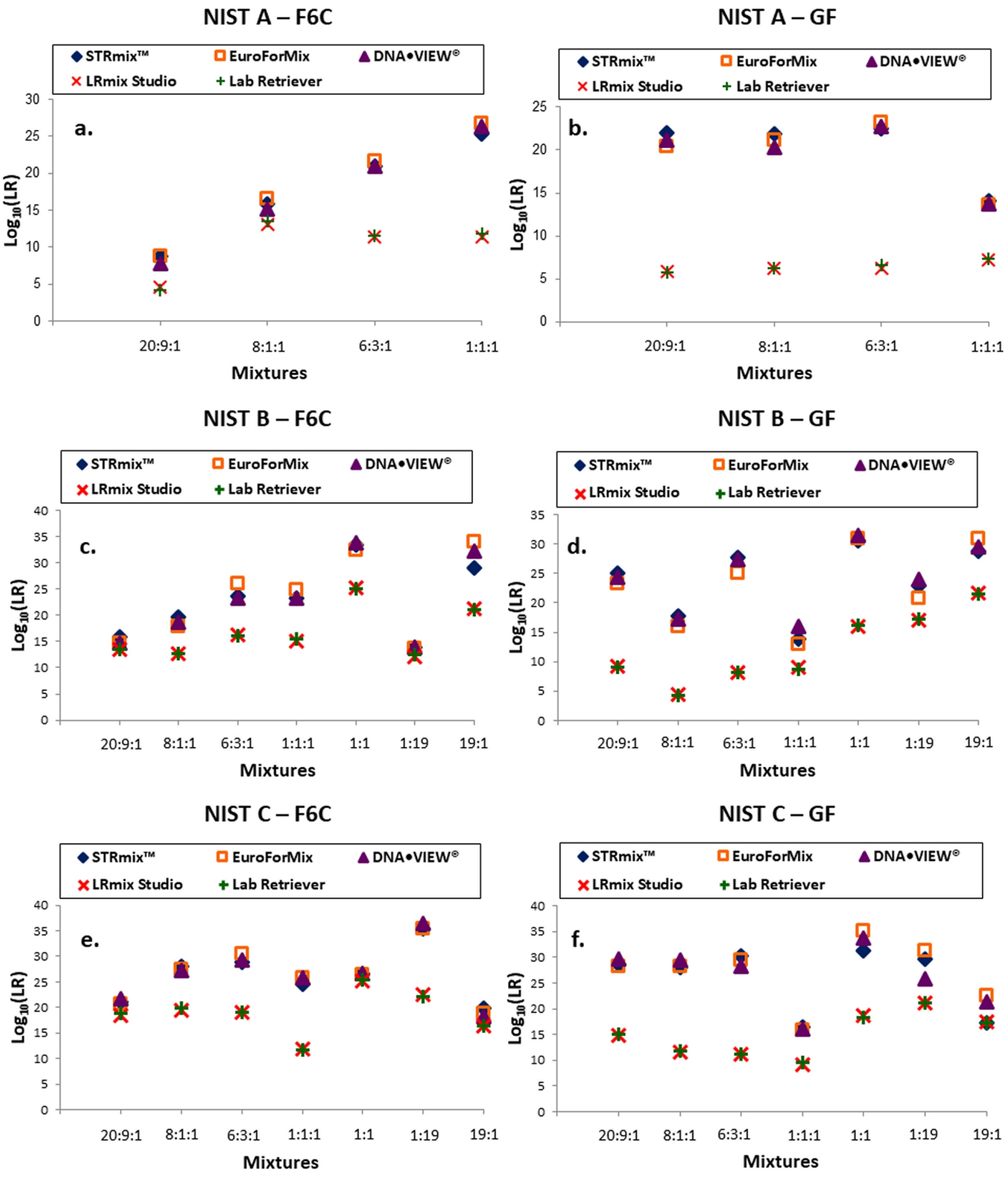
593

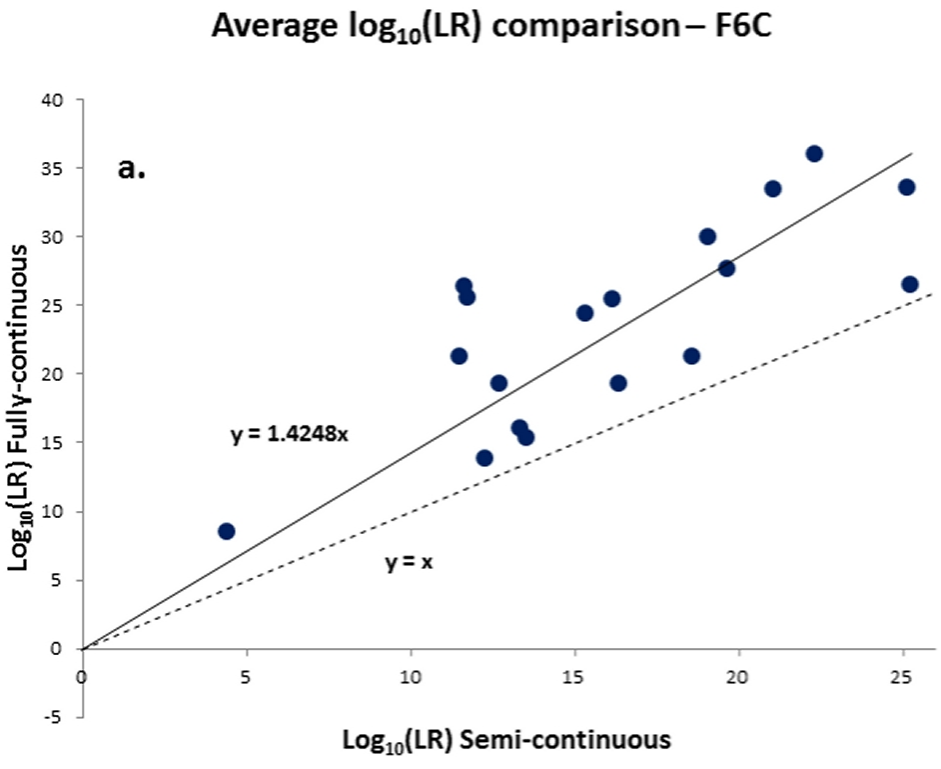
594

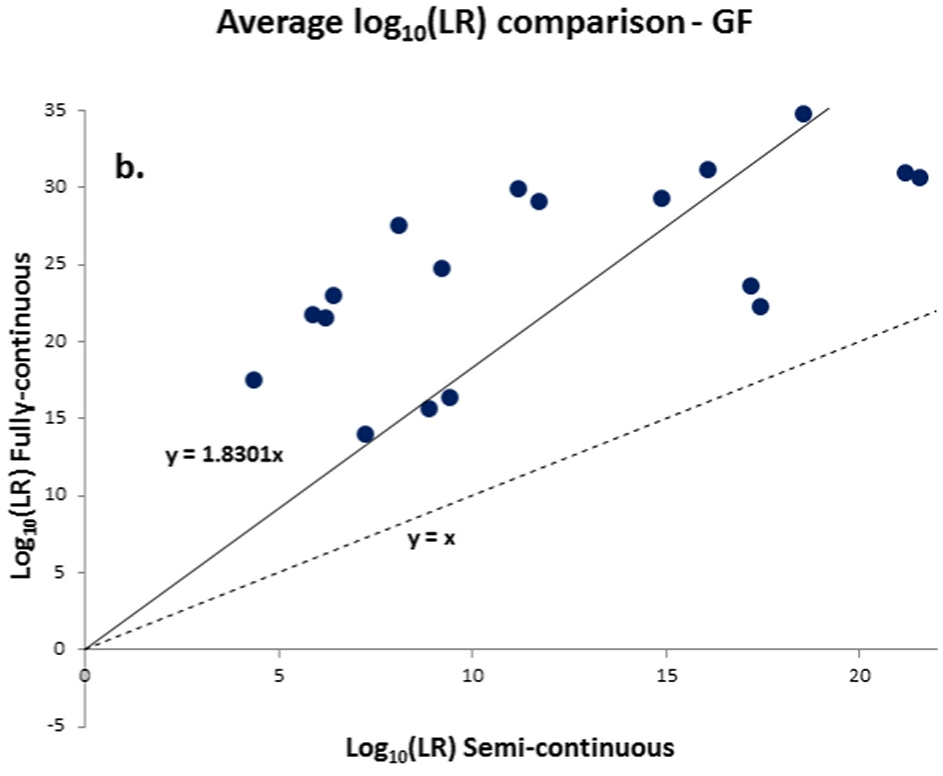
SD3 Channel Based Analytical Threshold					
Channel	DStdDev	Variance	Avg PBHW	Thresholds	SD3
Blue	5,002	25,022	10,000	47,455	48
Green	9,435	89,012	10,000	89,505	90
Yellow	4,977	24,766	10,000	47,212	48
Red	8,342	69,596	10,000	79,143	80
Purple	8,973	80,508	10,000	85,122	86
* Blue, Green, Yellow, Red, Purple channel(s) have no peaks.					

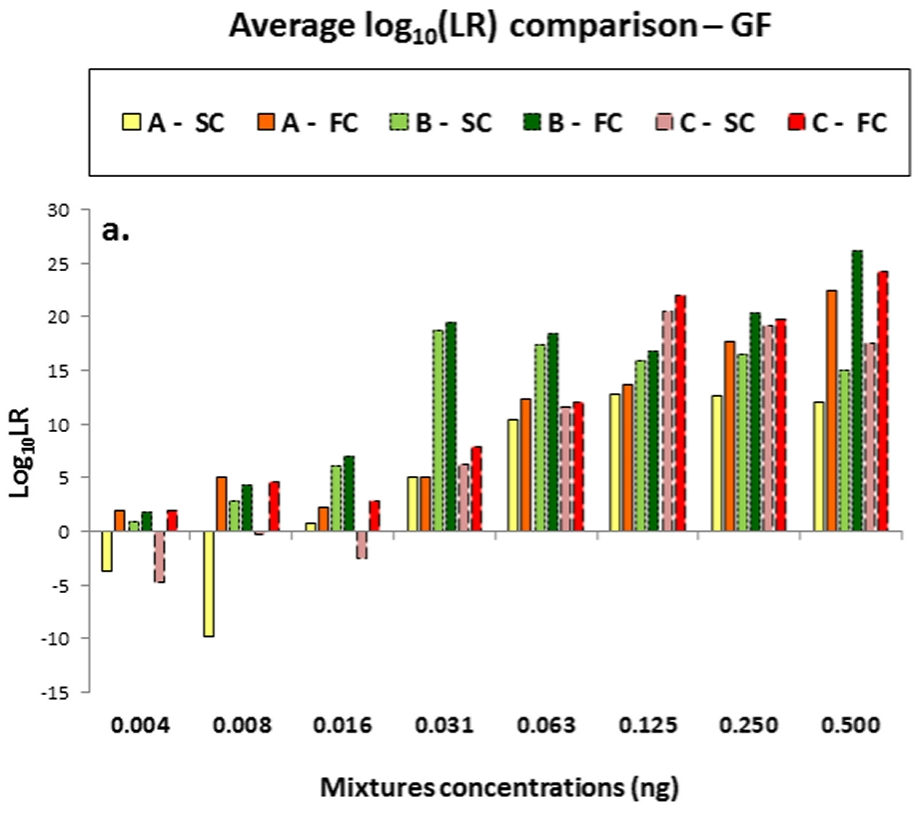
Channel Based RFU Extremums					
Channel	Max-Min Range	Avg Min RFU	Avg Max RFU	Avg RFU Range	Min-Max Fitted Lines Range Avg
Blue	60	-9,65	8,42	36,14	35,95
Green	132	-17,33	16,90	68,47	68,30
Yellow	76	-10,60	8,55	38,30	38,30
Red	96	-15,25	14,53	59,55	59,54
Purple	104	-16,60	16,84	66,88	66,86
Avg	93,60	-13,89	13,05	53,87	53,79

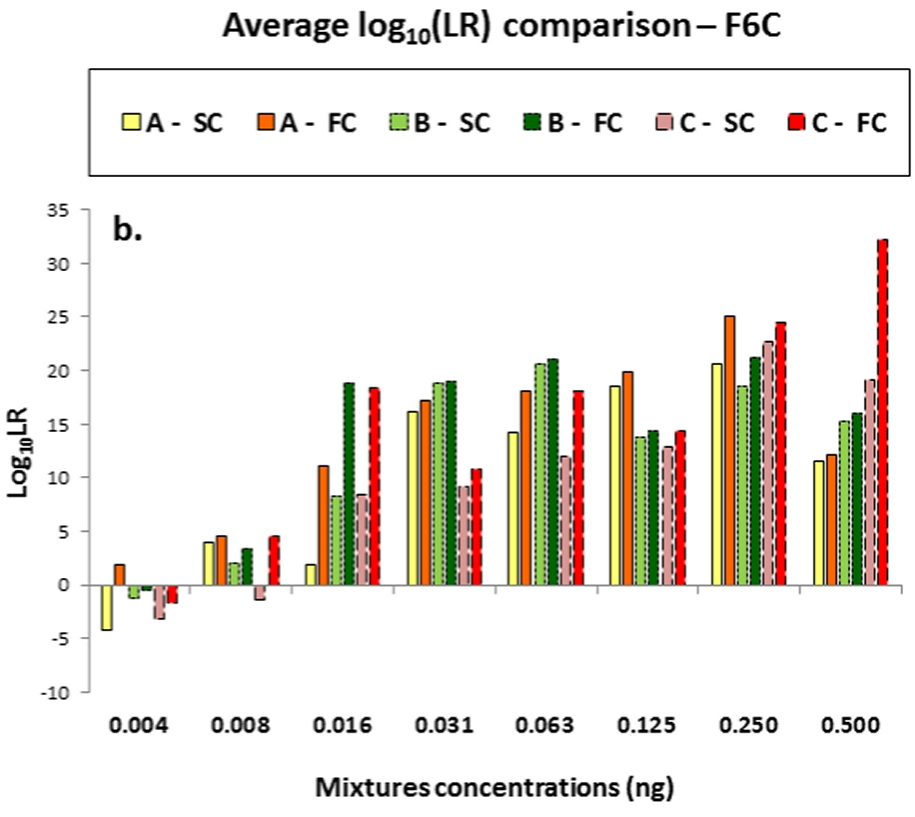
Figure 4 ArmedXpert™ results relative to the algorithms performing the calculation of differential analytical thresholds (i.e. per dye channel).


NIST A – F6C
STRmix™
EuroForMix
DNA•VIEW©
LRmix Studio
Lab Retriever
a.
Log10(LR)
20:9:1
8:1:1
6:3:1
1:1:1
Mixtures
NIST A – GF
b.
NIST B – F6C
c.
20:9:1
8:1:1
6:3:1
1:1:1
1:1
1:19
19:1
NIST B – GF
d.
NIST C – F6C
e.
NIST C – GF
f.


Average log10(LR) comparison – F6C
a.
y = 1.4248x
y = x
Log10(LR) Fully-continuous
Log10(LR) Semi-continuous
40
35
30
25
20
15
10
5
0
-5
0
5
10
15
20
25


Average log10(LR) comparison - GF
b.
y = 1.8301x
y = x
Log10(LR) Fully-continuous
Log10(LR) Semi-continuous
35
30
25
20
15
10
5
0
-5
0
5
10
15
20


Average log10(LR) comparison – GF
A - SC
A - FC
B - SC
B - FC
C - SC
C - FC
a.
Log10LR
30
25
20
15
10
5
0
-5
-10
-15
0.004
0.008
0.016
0.031
0.063
0.125
0.250
0.500
Mixtures concentrations (ng)


Average log10(LR) comparison – F6C
A - SC
A - FC
B - SC
B - FC
C - SC
C - FC
b.
Log10LR
35
30
25
20
15
10
5
0
-5
-10
0.004
0.008
0.016
0.031
0.063
0.125
0.250
0.500
Mixtures concentrations (ng)

SD3 Channel Based Analytical Threshold					
Channel	DStdDev	Variance	Avg PBHW	Thresholds	SD3
Blue	5,002	25,022	10,000	47,455	48
Green	9,435	89,012	10,000	89,505	90
Yellow	4,977	24,766	10,000	47,212	48
Red	8,342	69,596	10,000	79,143	80
Purple	8,973	80,508	10,000	85,122	86
* Blue, Green, Yellow, Red, Purple channel(s) have no peaks.					

Channel Based RFU Extremums					
Channel	Max-Min Range	Avg Min RFU	Avg Max RFU	Avg RFU Range	Min-Max Fitted Lines Range Avg
Blue	60	-9,65	8,42	36,14	35,95
Green	132	-17,33	16,90	68,47	68,30
Yellow	76	-10,60	8,55	38,30	38,30
Red	96	-15,25	14,53	59,55	59,54
Purple	104	-16,60	16,84	66,88	66,86
Avg	93,60	-13,89	13,05	53,87	53,79

Table 1 List of NIST samples used as known contributors for the preparation of DNA mixtures. Different proportions were evaluated for both 2- and 3-person mixtures. Referring to 2-person mixtures, different NIST materials were employed for the preparation of the samples according to the utilised DNA amplification kit.

Known 2-person mixtures		
DNA Typing Kit	**Reference Material**	**Mixtures ratios (0.500 ng)**
ESI 17 Fast	B:C / male:male	19:1 8:1 1:1 1:19
ESX 17 Fast	A:C / female:male	
Fusion	B:C / male:male	
Fusion 6C	B:C / male:male	
GlobalFiler	B:C / male:male	
Mini Filer	A:C / female:male	
NGM SElect	A:C / female:male	

Known 3-person mixtures		
DNA Typing Kit	**Reference Material**	**Mixtures ratios (0.500 ng)**
All kits	A:B:C / Female:male:male	1:1:1 6:3:1 8:1:1 20:9:1

Table 2 Log(LR) results relative to the interpretation process performed on the DNA mixture collected on a cap recovered on a crime scene. 3SD and Min-Max represent the algorithms that were employed to evaluate the differential analytical thresholds. POI represents the suspect (i.e. the person of interest), while U stands for unknown(s) individual(s) extracted from the allele frequencies reference dataset [47].

Analytical Threshold	**3SD**	**Min-Max**	**3SD**	**Min-Max**
Hp	S + 1U	S + 1U	S + 2U	S + 2U
Hd	2U	2U	3U	3U
Software			**Log(LR)**	
LRmix Studio	-0.37	-0.16	2.36	2.48
Lab Retriever	-0.36	-0.16	2.37	2.49
DNA•VIEW®	2.29	3.47	6.79	7.06
EuroForMix	2.33	3.60	6.25	7.12
STRmix™	2.84	3.63	7.01	7.03
Interpretative decision	**Inconclusive**	**Inconclusive**	**Support to H(p)**	**Support to H(p)**

Authors' contributions

EA conceived the study, carried out the multivariate studies and drafted the manuscript.

MO, GD, DC and PG participated in the design of the study, carried out the genotyping step and helped to draft the manuscript.

MV participated in the design of the study and helped to draft the manuscript.

All authors read and approved the final manuscript.

Acknowledgements

This study was partially founded by the Fondazione Giovanni Goria, within the grant program named “Bando Talenti della Società Civile”. Continuous support from M.I.U.R. and Regione Piemonte is kindly acknowledged.

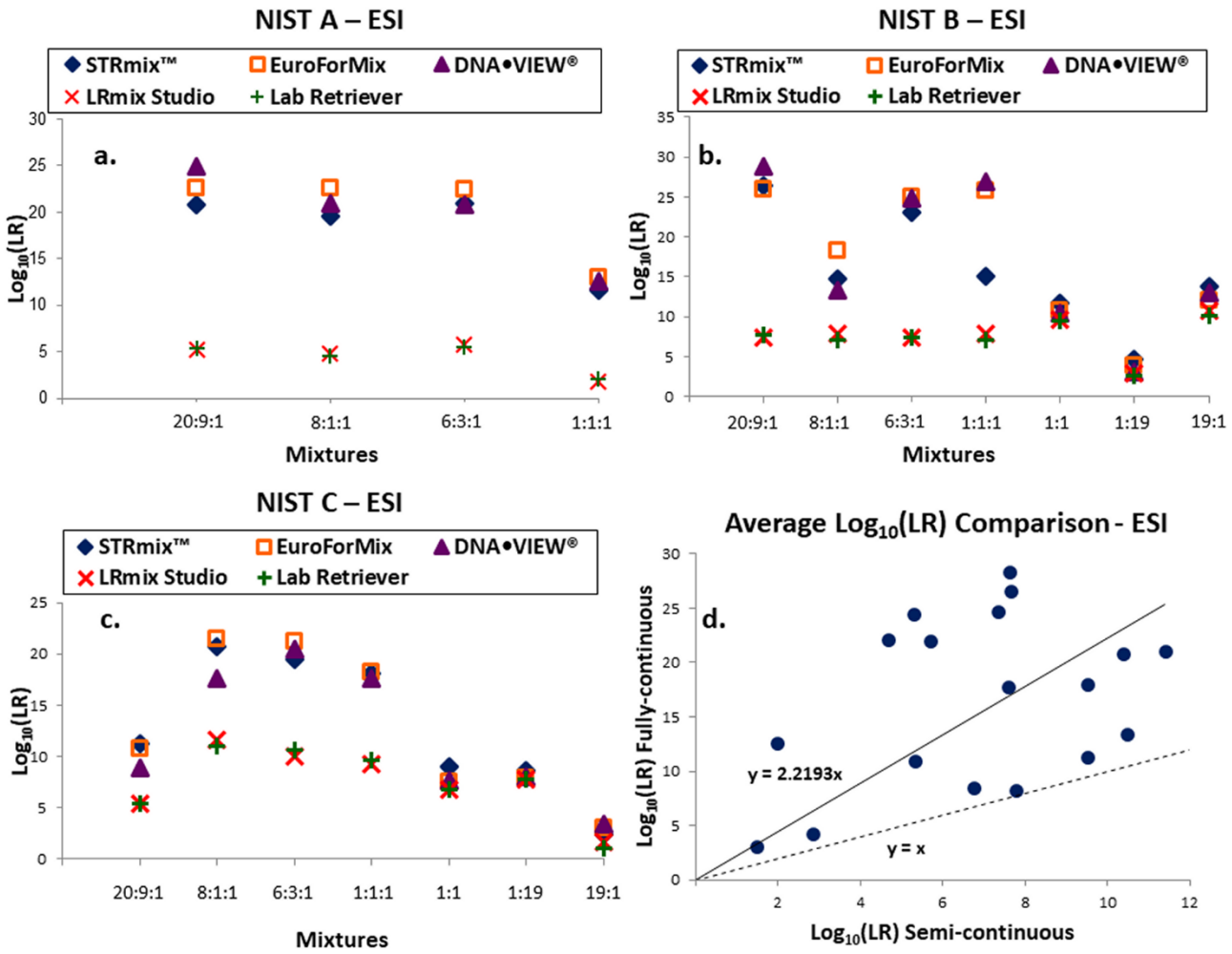


Figure S1 Log(LR) results relative to 2-person and 3-person mixtures that were amplified with ESI DNA amplification kit. Log(LR) results of the known contributor labelled as NIST A (a), NIST B (b) and NIST C (c) are shown. In particular, log(LR) values are represented by: blue diamonds for STRmix™, orange squares for Euroformix, purple triangles for DNA•VIEW®, red addition marks for LRmix Studio and green crosses for Lab Retriever. Average log(LR) values provided by semi-continuous and fully-continuous models for ESI DNA amplification kit are remarked in (d). The dashed line represents an hypothetic situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.

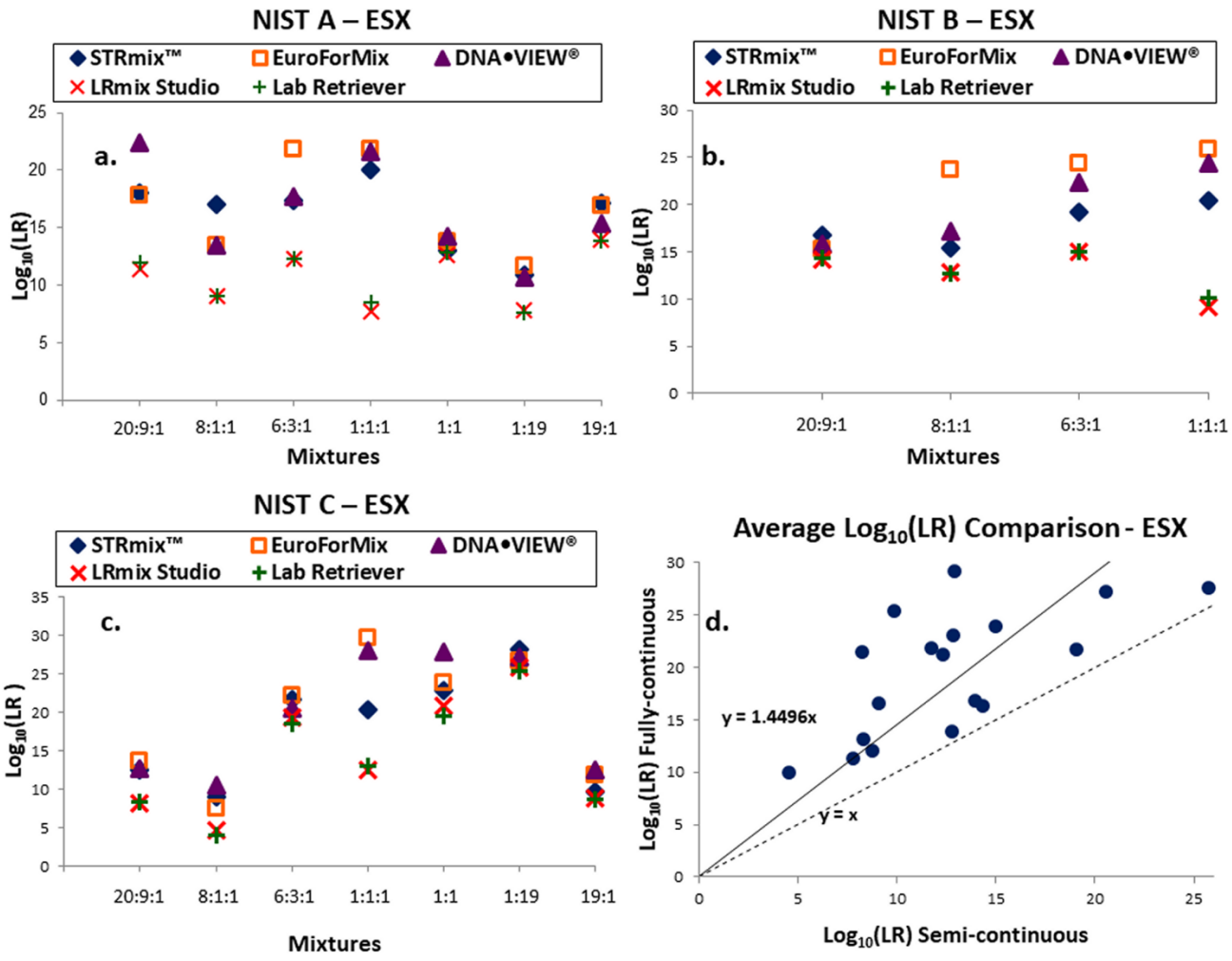


Figure S2 Log(LR) results relative to 2-person and 3-person mixtures that were amplified with ESX DNA amplification kit. Log(LR) results of the known contributor labelled as NIST A (a), NIST B (b) and NIST C (c) are shown. In particular, log(LR) values are represented by: blue diamonds for STRmix™, orange squares for Euroformix, purple triangles for DNA•VIEW®, red addition marks for LRmix Studio and green crosses for Lab Retriever. Average log(LR) values provided by semi-continuous and fully-continuous models for ESX DNA amplification kit are remarked in (d). The dashed line represents an hypothetic situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.

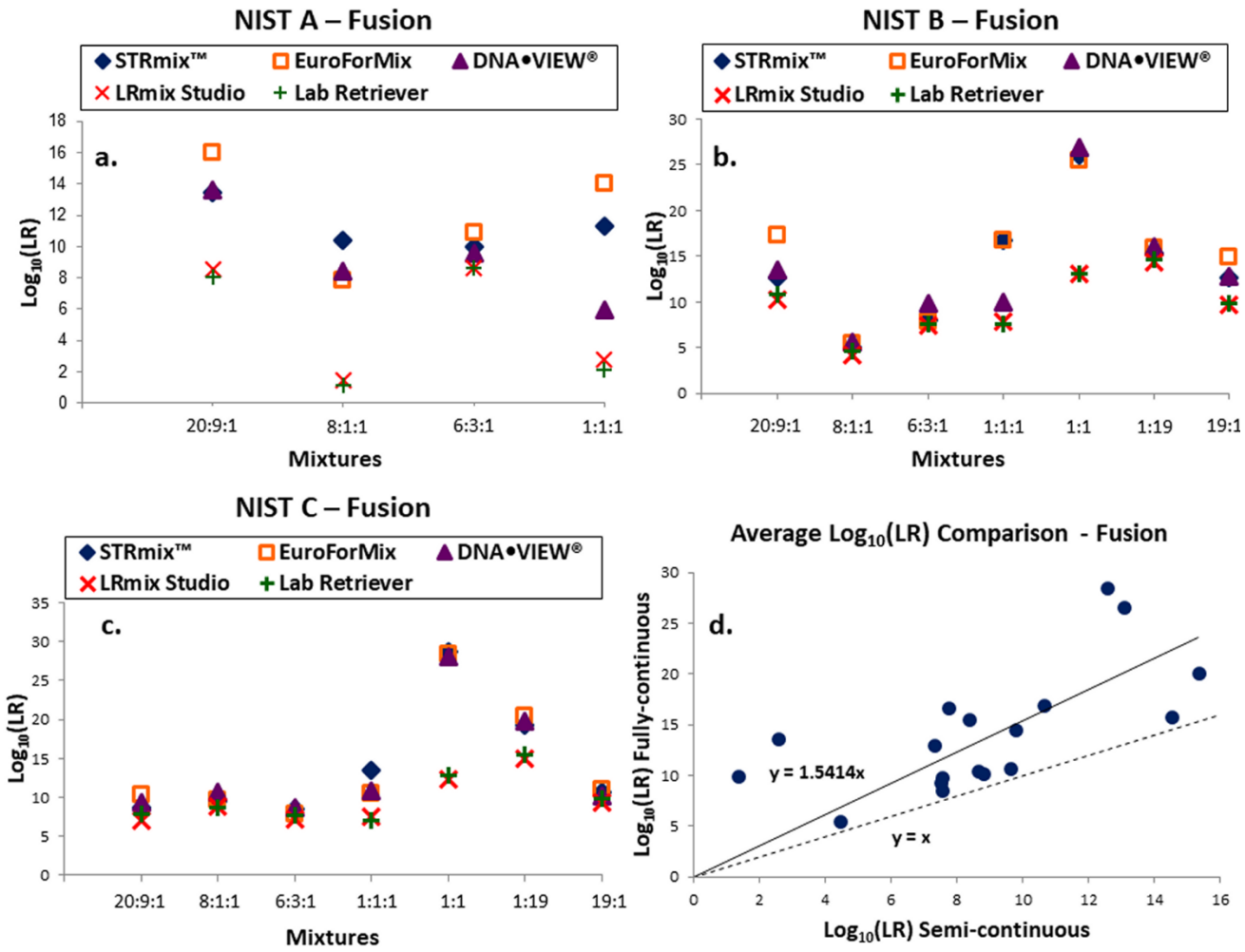


Figure S3 Log(LR) results relative to 2-person and 3-person mixtures that were amplified with Fusion DNA amplification kit. Log(LR) results of the known contributor labelled as NIST A (a), NIST B (b) and NIST C (c) are shown. In particular, log(LR) values are represented by: blue diamonds for STRmix™, orange squares for Euroformix, purple triangles for DNA•VIEW®, red addition marks for LRmix Studio and green crosses for Lab Retriever. Average log(LR) values provided by semi-continuous and fully-continuous models for Fusion DNA amplification kit are remarked in (d). The dashed line represents an hypothetic situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.

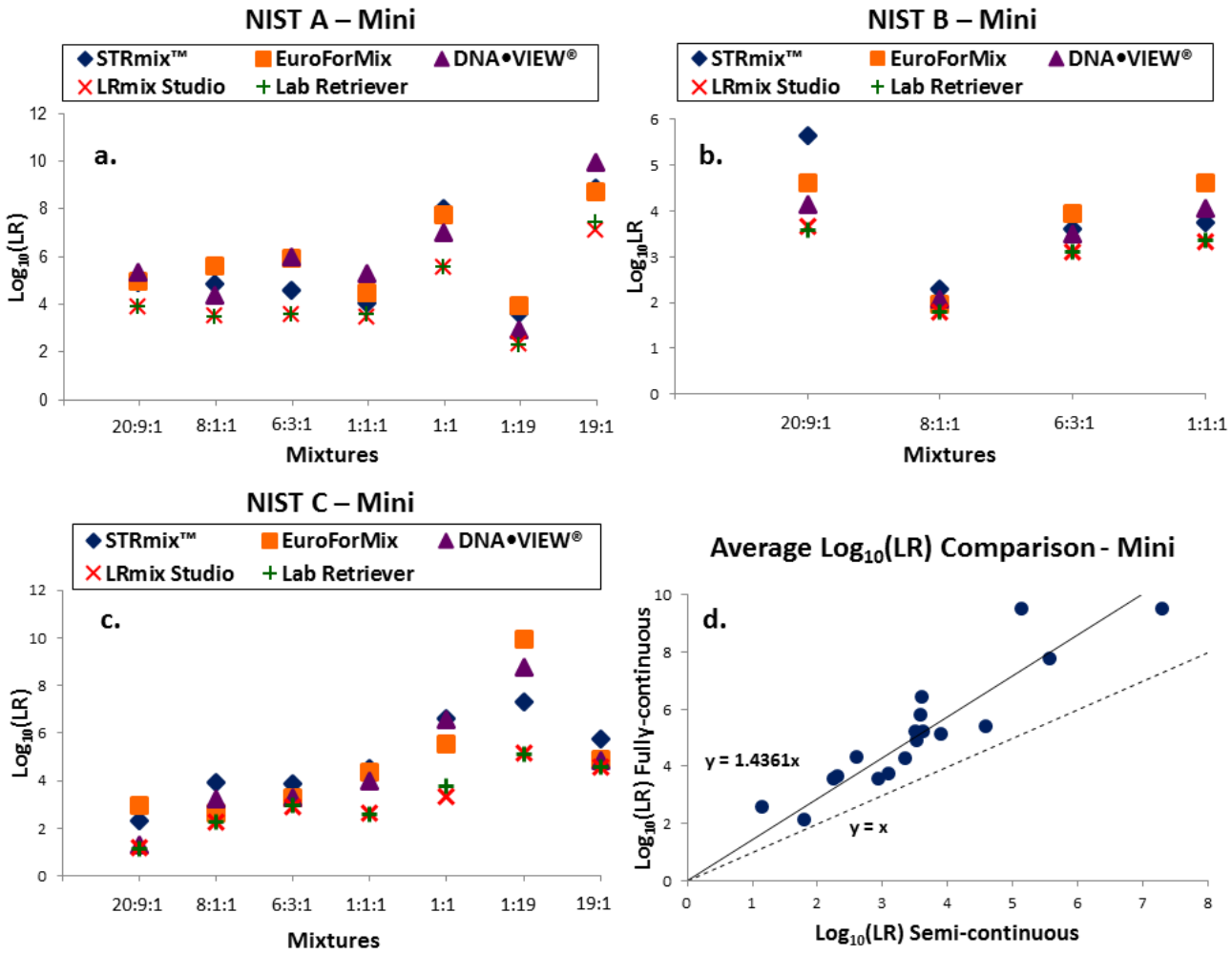


Figure S4 Log(LR) results relative to 2-person and 3-person mixtures that were amplified with Mini DNA amplification kit. Log(LR) results of the known contributor labelled as NIST A (a), NIST B (b) and NIST C (c) are shown. In particular, log(LR) values are represented by: blue diamonds for STRmix™, orange squares for Euroformix, purple triangles for DNA•VIEW®, red addition marks for LRmix Studio and green crosses for Lab Retriever. Average log(LR) values provided by semi-continuous and fully-continuous models for Mini DNA amplification kit are remarked in (d). The dashed line represents an hypothetic situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.

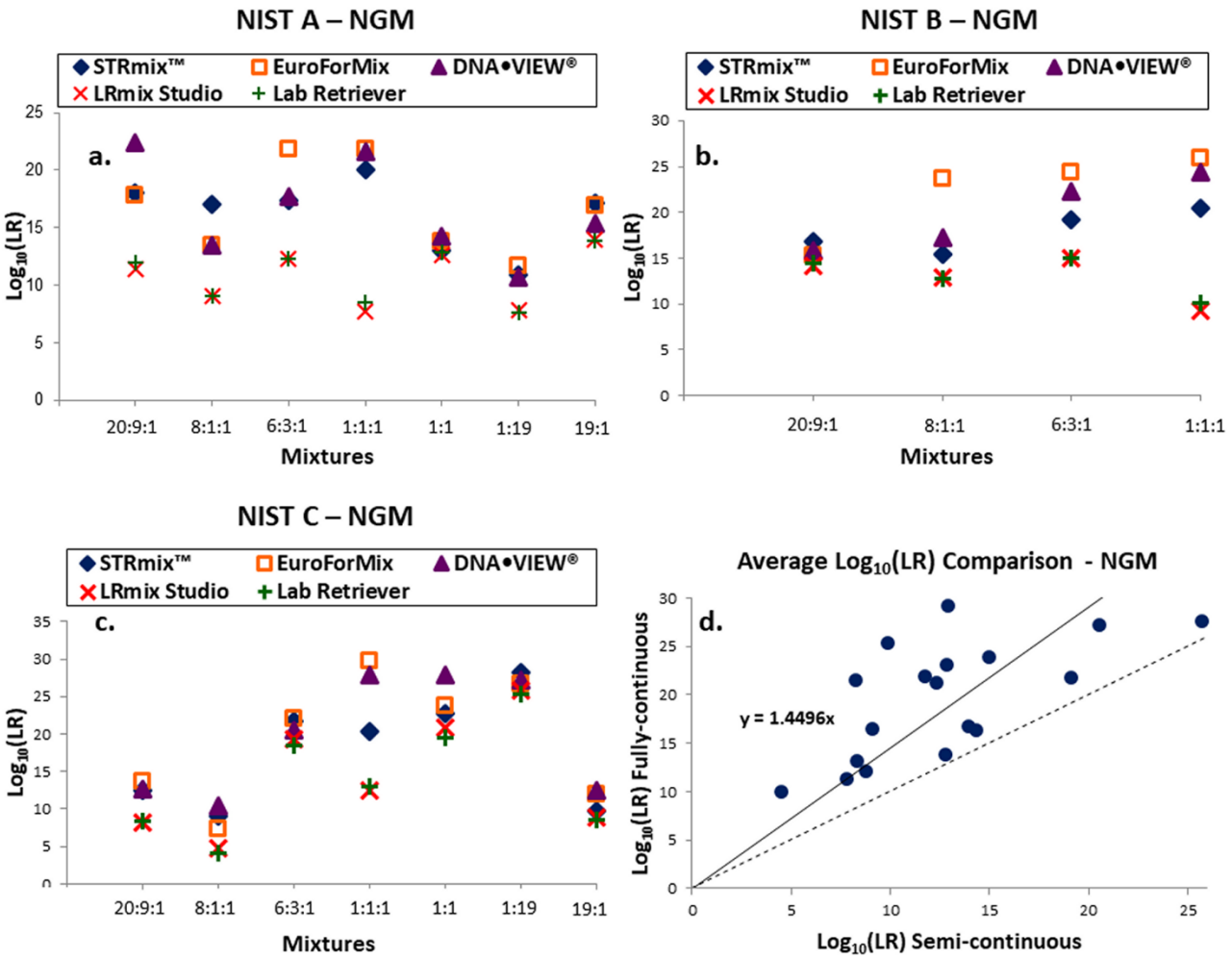


Figure S5 Log(LR) results relative to 2-person and 3-person mixtures that were amplified with NGM DNA amplification kit. Log(LR) results of the known contributor labelled as NIST A (a), NIST B (b) and NIST C (c) are shown. In particular, log(LR) values are represented by: blue diamonds for STRmix™, orange squares for Euroformix, purple triangles for DNA•VIEW®, red addition marks for LRmix Studio and green crosses for Lab Retriever. Average log(LR) values provided by semi-continuous and fully-continuous models for NGM DNA amplification kit are remarked in (d). The dashed line represents an hypothetic situation where all the log(LR) results are the same for both the investigated models, while the solid line (with intercept equal to zero) indicates the average trend that is observed among the results provided by semi- and fully-continuous algorithms.

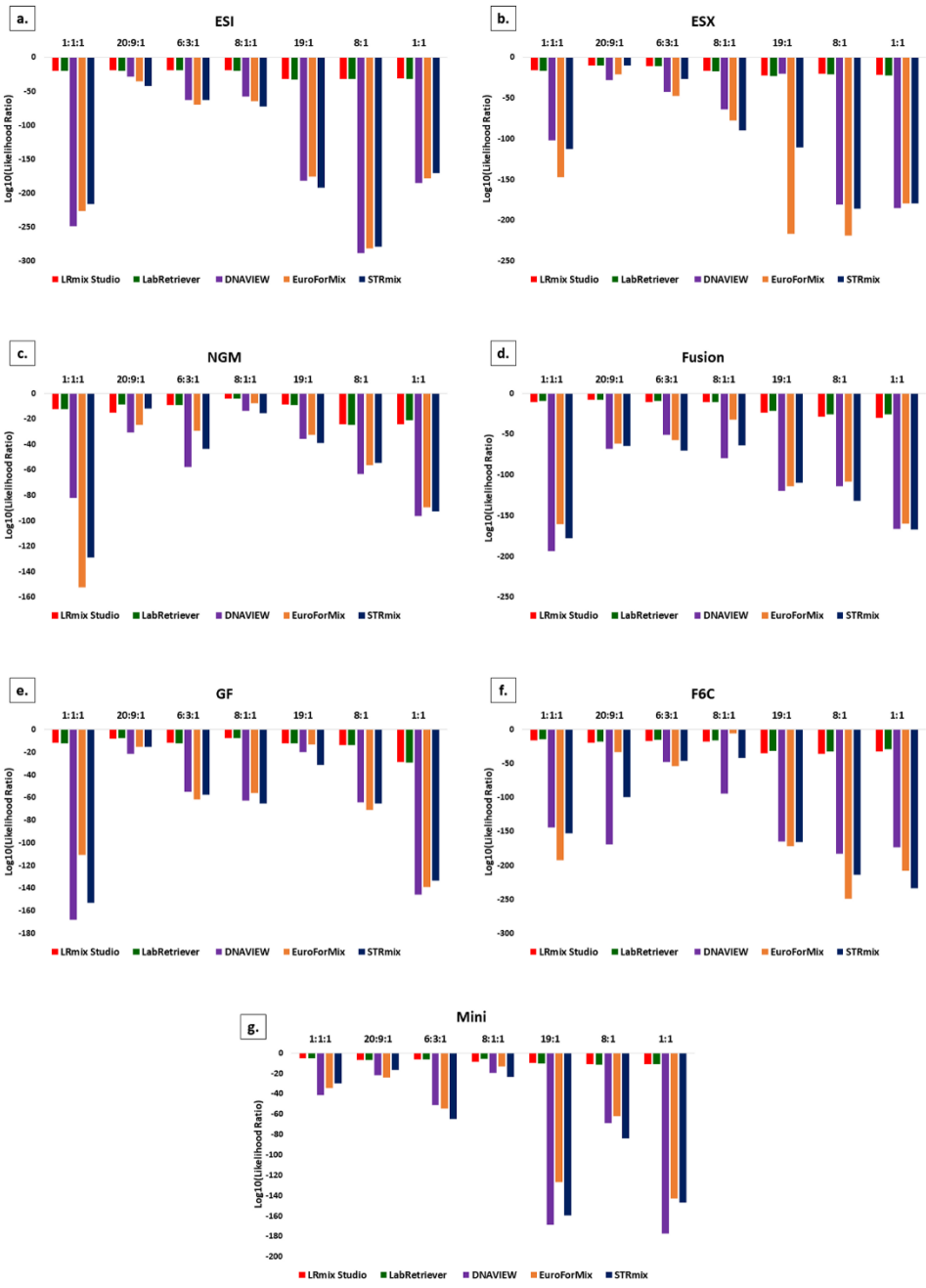


Figure S6 Histograms displaying false donor (i.e. NIST F) log(LR) values log(LR) results relative to 2-person and 3-person mixtures that were amplified with different DNA amplification kits. DNA mixtures were amplified by ESI (a), ESX (b), NGM (c), Fusion (d), GlobalFiler (e), Fusion 6C (f) and MiniFiler (g) DNA amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. Log(LR) values are represented by red (LRmix Studio), green (Lab Retriever), purple (DNAVIEW®), orange (EuroForMix) and blue (STRmix™) histograms.

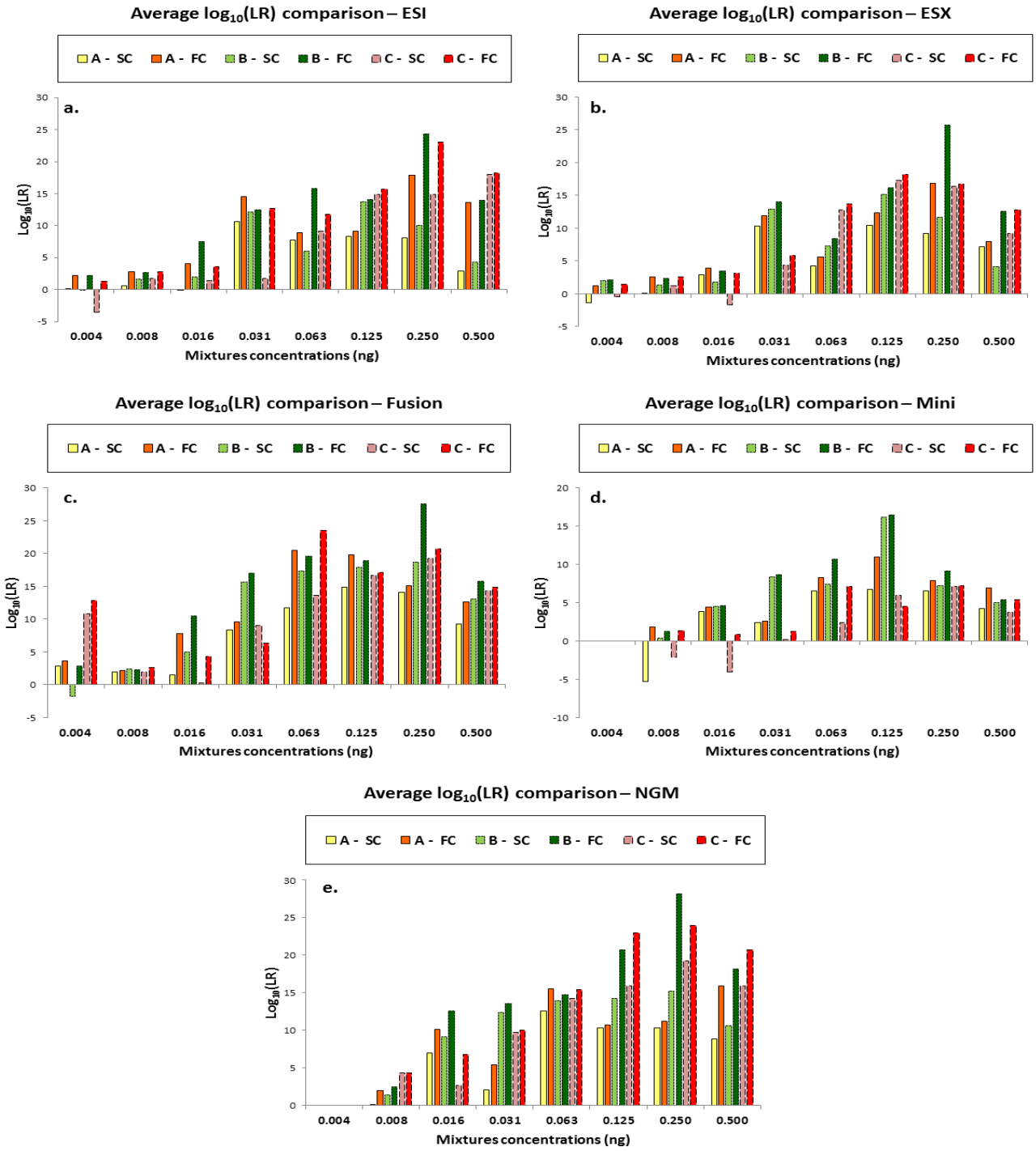


Figure S7 Histograms displaying average log(LR) values provided by semi-continuous (SC) and fully-continuous (FC) models for 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified by ESI (a), ESX (b), Fusion (c), Mini (d) and NGM (c) DNA amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. Average log(LR) values relative to NIST materials composing the mixtures are represented by yellow (SC) and orange (FC) histograms for NIST A contributor, light green (SC) and dark green (FC) histograms for NIST B contributor, pink (SC) and red (FC) histograms for NIST C contributor. Data relevant to 0.004 ng mixtures for Mini and NGM are not available as such DNA mixtures didn't provide any accessible results when amplified by such DNA amplification kits.

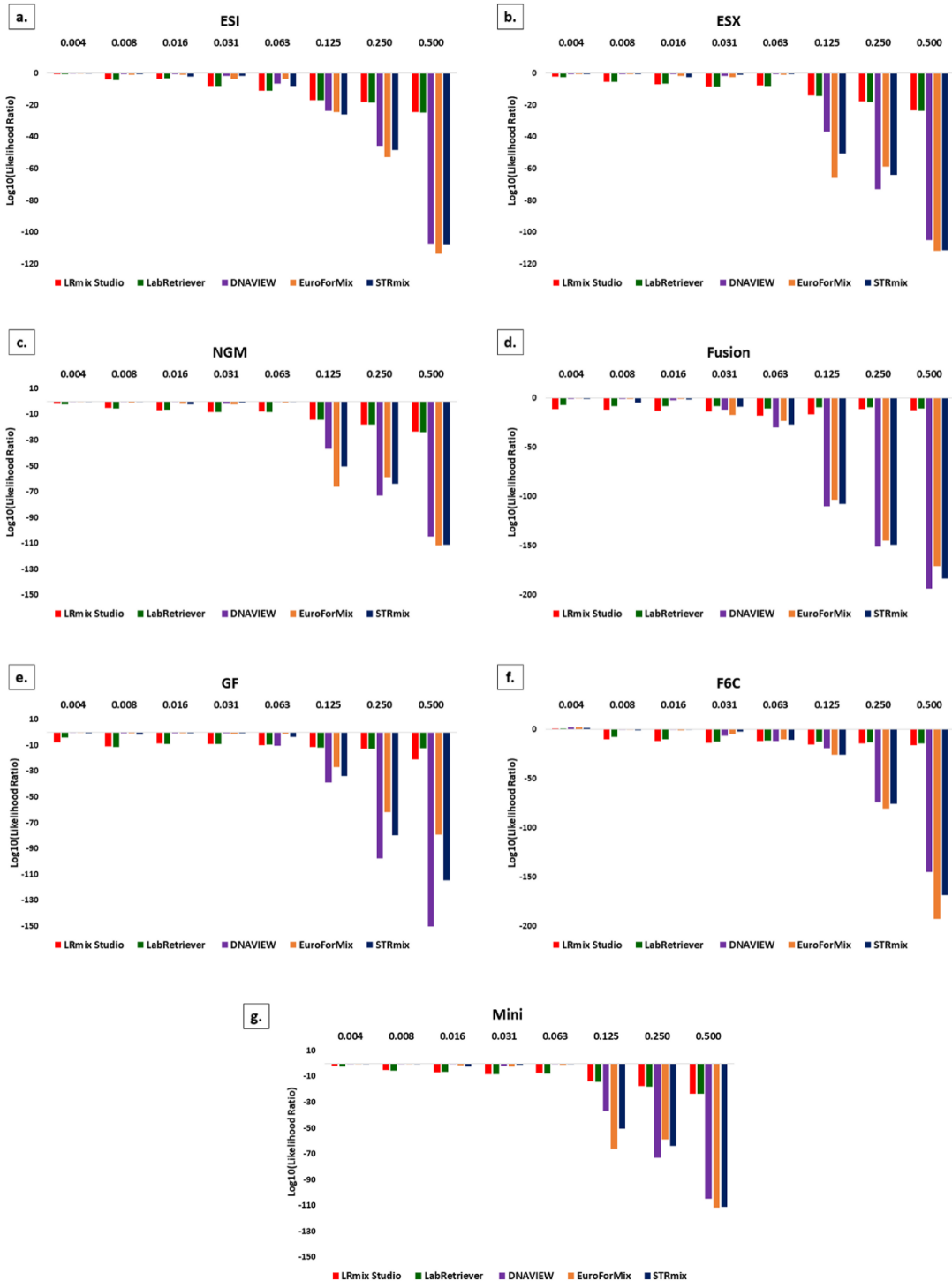


Figure S8 Histograms displaying false donor (i.e. NIST F) log(LR) values log(LR) results relative to the 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified with different DNA amplification kits. DNA mixtures were amplified by ESI (a), ESX (b), NGM (c), Fusion (d), GlobalFiler (e), Fusion 6C (f) and MiniFiler (g) DNA amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. Log(LR) values are represented by red (LRmix Studio), green (Lab Retriever), purple (DNAVIEW®), orange (EuroForMix) and blue (STRmix™) histograms.

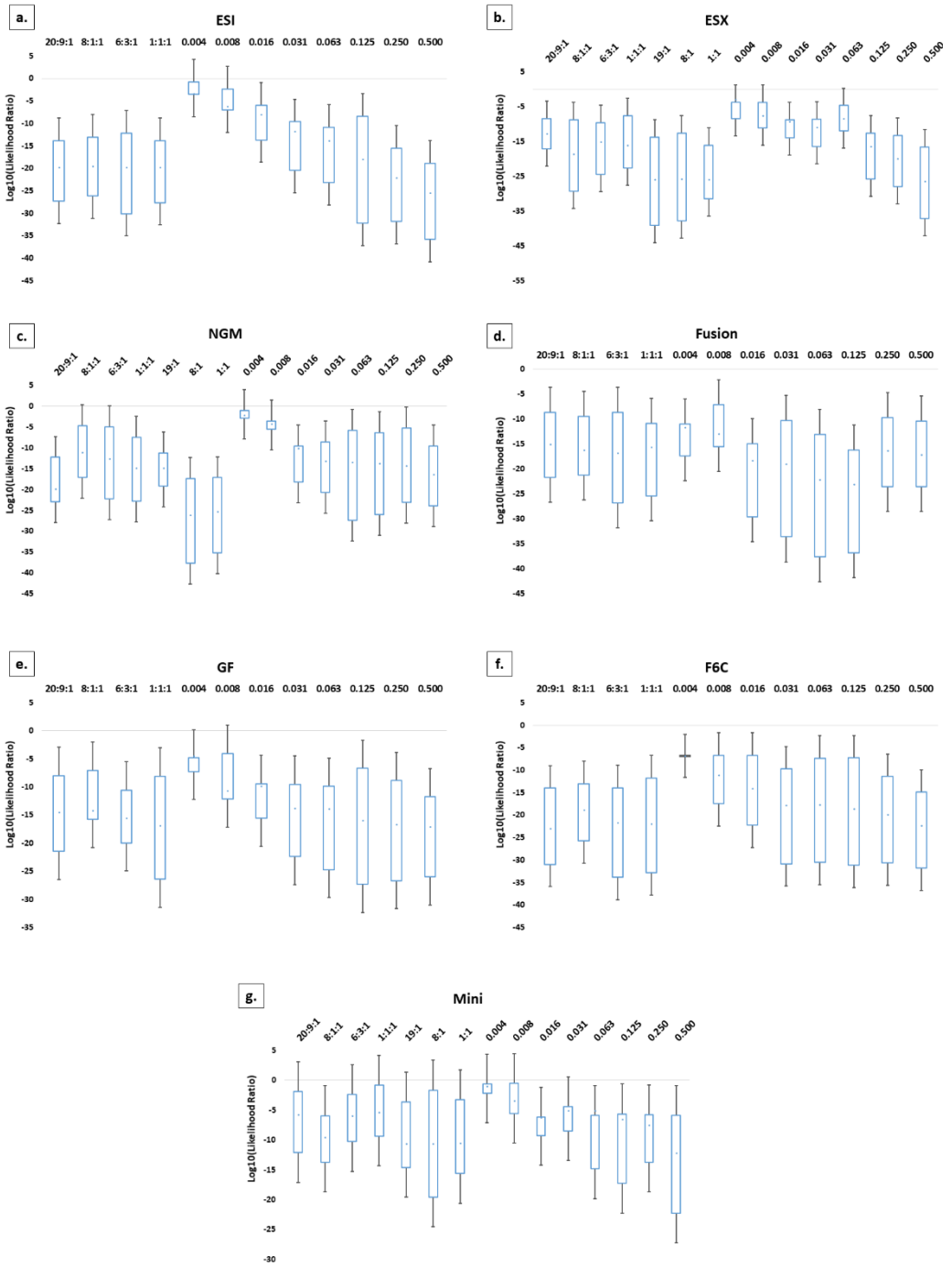


Figure S9a Boxplots displaying non-contributor tests for NIST A and their relative log(LR) values relative to the 2-person and 3-person mixtures and the 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified with different DNA amplification kits. DNA mixtures were amplified by ESI (a), ESX (b), NGM (c), Fusion (d), GlobalFiler (e), Fusion 6C (f) and MiniFiler (g) DNA amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. The dot within the boxplot represents the median value.

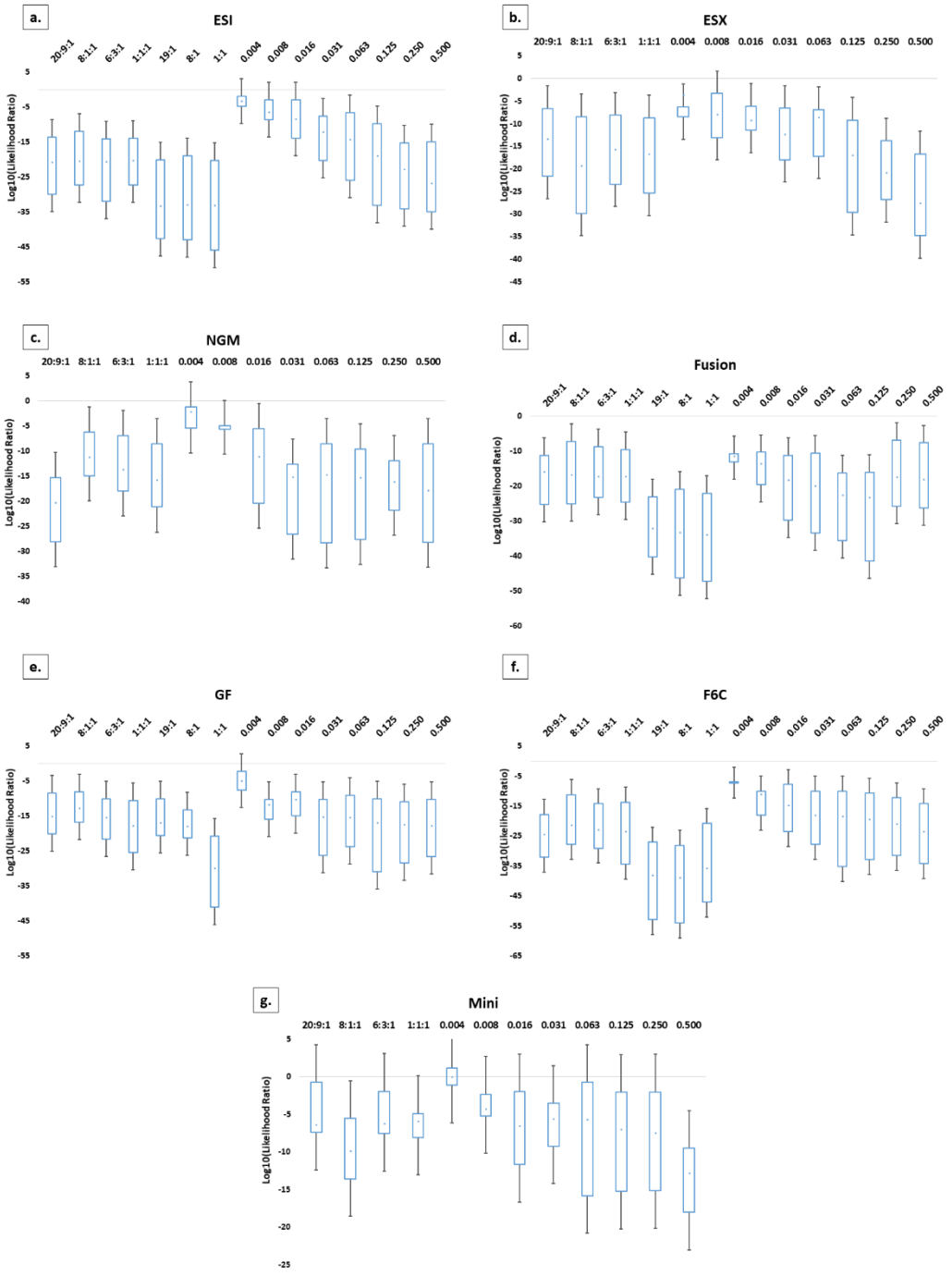


Figure S9b Boxplots displaying non-contributor tests for NIST B and their relative log(LR) values relative to the 2-person and 3-person mixtures and the 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified with different DNA amplification kits. DNA mixtures were amplified by ESI (a), ESX (b), NGM (c), Fusion (d), GlobalFiler (e), Fusion 6C (f) and MiniFiler (g) DNA amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. The dot within the boxplot represents the median value.

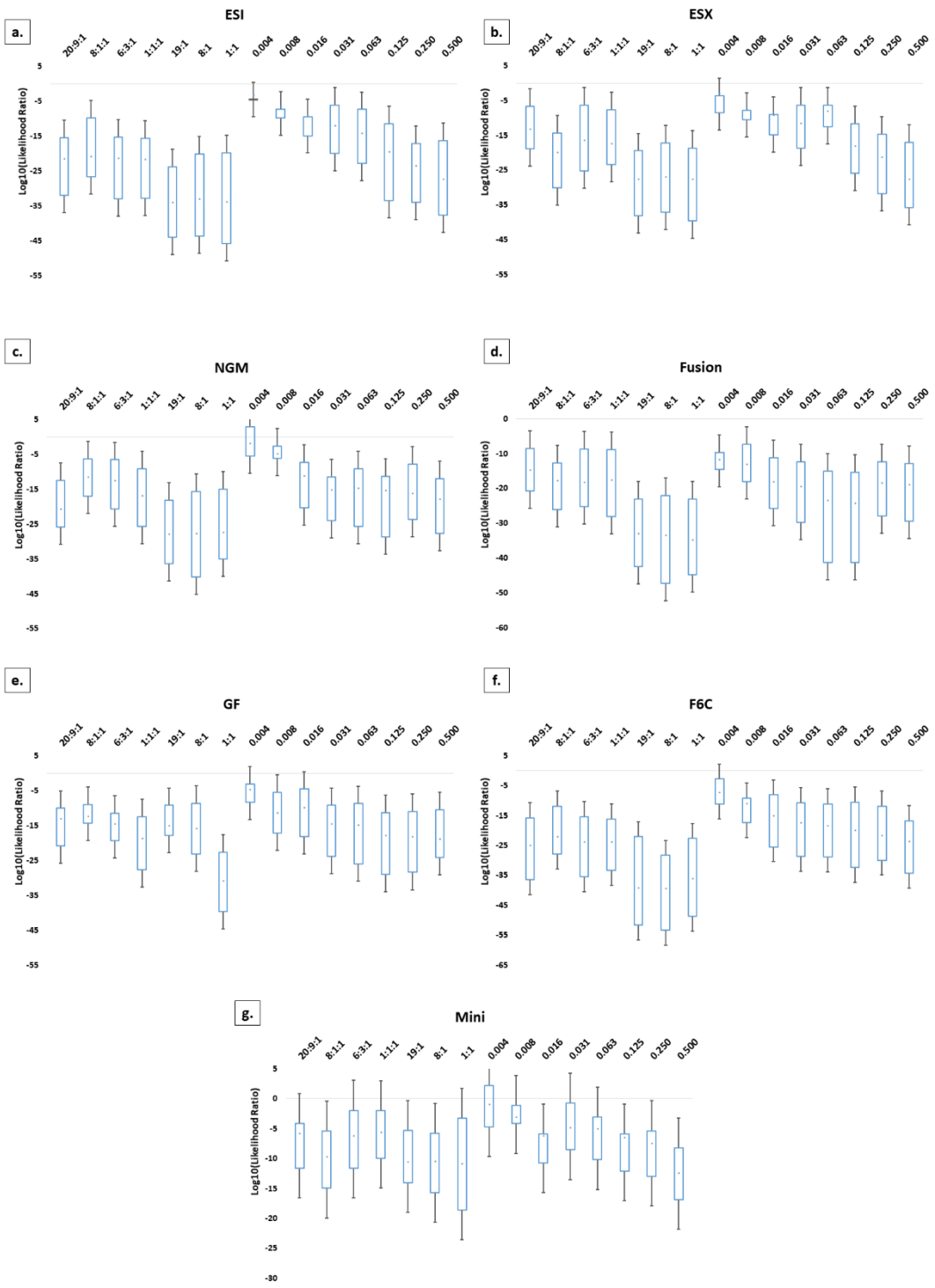


Figure S9c Boxplots displaying non-contributor tests for NIST C and their relative log(LR) values relative to the 2-person and 3-person mixtures and the 1:1:1 serially scalarly-diluted 3-person DNA mixtures that were amplified with different DNA amplification kits. DNA mixtures were amplified by ESI (a), ESX (b), NGM (c), Fusion (d), GlobalFiler (e), Fusion 6C (f) and MiniFiler (g) DNA amplification kits. The codes indicating the different DNA mixtures are reported on the x axis. The dot within the boxplot represents the median value.